DOCUMENT RESUME

ED 211 609 TM 820 097

AUTHOR Haebara, Tomokazu
TITLE Least Squares Method for Equating Logistic Ability Scales: A General Approach and Evaluation. Iowa Testing Programs Occasional Papers, Number 30.
INSTITUTION Iowa Univ., Iowa City.
PUB DATE Dec 81
NOTE 116p.; Ph.D. Dissertation, University of Iowa.

EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS *Academic Ability; *Equated Scores; *Latent Trait Theory; *Least Squares Statistics; Mathematical Models; *Scaling; Test Construction; Test Items
IDENTIFIERS Computer Simulation

ABSTRACT

When several ability scales in item response models are separately derived from different test forms administered to different samples of examinees, these scales must be equated to a common scale because their units and origins are arbitrarily determined and generally different from scale to scale. A general method for equating logistic ability scales was developed that uses a least squares method with weights reflecting the relative importance of correct equating at the various values on the ability continuum. The method is characterized by explicit definition of the equating error on the probability scale, minimization of the overall error, and coherence of the equating in all directions. It is applicable when the test forms are linked by common items, and various linkage patterns can be used. When there are more than m - 1 links among m test forms, a restricted solution is provided to insure an internally consistent system of equatings. Results of the application of the method to many computer-generated data sets indicated that the accuracy of the obtained estimates of the equating constants was generally satisfactory and that the number of common items in the linked pairs of test forms was the most important factor in determining the accuracy of estimation. (Author/GK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED211609

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

IOWA TESTING PROGRAMS OCCASIONAL PAPERS

Number 30 - December 1981

Least Squares Method for Equating Logistic Ability Scales: A General Approach and Evaluation

Tomokazu Haebara

Iowa Testing Programs

The University of Iowa

TM 820 097

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
T. Haebara
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ABSTRACT

When several ability scales in item response models are separately derived from different test forms administered to different samples of examinees, these scales must be equated to a common scale because their units and origins are arbitrarily determined and generally different from scale to scale. In this paper, a general method for equating logistic ability scales was developed that uses a least squares method with weights reflecting the relative importance of correct equating at the various values on the ability continuum. The method is characterized by the explicit definition of the equating error on the probability scale, the minimization of the overall error, and the coherence of the equating in all directions. It is applicable when the test forms are linked by common items, and various linkage patterns can be used. When there are more than m - 1 links among m test forms, a restricted solution is provided to insure an internally consistent system of equatings. The results of the application of the method to many computer-generated data sets indicated that the accuracy of the obtained estimates of the equating constants was generally satisfactory and that the number of common items in the linked pairs of test forms was the most important factor in determining the accuracy of estimation.

ACKNOWLEDGMENTS

This paper has been approved by the examining committee as the thesis for the Ph.D. degree in Education at the University of Iowa (December 1981 graduation). I wish to thank the chairman of the committee, Dr. William E. Coffman, and the following members of the committee for their valuable contributions to the study: Dr. Leonard S. Feldt, Dr. Robert A. Forsyth, Dr. Robert V. Hogg, Dr. Hiram D. Hoover, and Dr. Melvin R. Novick. I am also indebted to Ms. Jan Lewis and Mr. Shin-ichi Mayekawa for their helpful comments on the earlier versions of the paper.

TABLE OF CONTENTS

	Page
LIST OF TABLES	v
LIST OF FIGURES	vii
PART A. PROBLEM	1
CHAPTER	
I. INTRODUCTION	2
Item Response Theory in Mental Test Theory	2
Ability Scale in Item Response Theory	5
Equating of Ability Scales	6
Purpose of the Study	8
II. FORMULATION OF THE PROBLEM	10
Logistic Item Response Model	10
Equating Problem and Some Remarks	13
Equating Designs	18
Previously Developed Methods	20
PART B. THE METHOD	26
CHAPTER	
III. UNRESTRICTED CASE	27
Least Squares Criterion	27
Minimization of Criterion Function	31
Remarks on Weight Functions	33
Related Methods	37
Equating More Than Two Scales	40
Common Item Response Functions	41
IV. GENERALIZATION OF THE METHOD	44
Another Approach to the Unrestricted Case	44

Page

General Method 47
Algorithm 56
Index of Internal Consistency 58

PART C. NUMERICAL EVALUATION 60

CHAPTER

V. EVALUATION DESIGNS 61

Overview 61
Evaluation 1 63
Evaluation 2 71
Evaluation 3 72
Standardized Criterion Function 77

VI. RESULTS AND DISCUSSION 78

Evaluation 1 78
Evaluation 2 84
Evaluation 3 88

VII. SUMMARY 97

General Equating Method 97
Numerical Evaluation of the Method 98

REFERENCES 100

LIST OF TABLES

Table		Page
1.	Approximate Standard Errors of Two Item Response Functions at Nine Ability Levels	36
2.	Organization of the ITBS, Test V	49
3.	Organization of the CAT, Nonverbal Battery, Test 3	49
4.	Means and SD's of Abilities in the 4 Examinee Samples and True Equating Constants	63
5.	True Item Parameters in the 4 Item Sets Defined on the Common Scale	65
6.	Organization of the 4 Test Forms and Their Allocation to Examinee Samples	69
7.	Means and SD's of Abilities in the 8 Examinee Samples and True Equating Constants	73
8.	Means and SD's of Abilities in the 12 Examinee Samples and True Equating Constants	74
9.	Ranges of Item Difficulties in the 12 Item Sets Defined on the Common Scale	75
10.	Means and SD's of the 100 Estimates of Equating Constants: Restricted Solution	79
11.	Means and SD's of the 100 Estimates of Equating Constants: Unrestricted Solution	80
12.	Oblique Procrustes Factor Solution for the Correlations among the Estimates: Restricted Solution	82
13.	Oblique Procrustes Factor Solution for the Correlations among the Estimates: Unrestricted Solution	83
14.	Various Comparisons Between the Restricted and the Unrestricted Solutions in Item Parameter Estimation	86

15. Means of Absolute Errors of Estimates of k's in 27 Closed Chains under 4 Conditions 90

16. Means of Absolute Errors of Estimates of d's in 27 Closed Chains under 4 Conditions 91

17. Minimized Values of Standardized Criterion Function in 27 Closed Chains under 4 Conditions 92

18. Means of the Values in Tables 15, 16, and 17 95

LIST OF FIGURES

Figure		Page
1.	Open chain of four test forms linked by common items.	19
2.	Closed chain of four test forms linked by common items.	20
3.	Incomplete network of four test forms linked by common items.	21
4.	Complete network of four test forms linked by common items.	21
5.	Incomplete network of 6 levels of the ITBS, Test V.	50
6.	Incomplete network of 8 levels of the CAT, Nonverbal Battery, Test 3.	50
7.	Two examples of unrestrictive design.	54

PART A

PROBLEM

CHAPTER I

INTRODUCTION

Item Response Theory in Mental Test Theory

Mental test theory, a major division of psychometrics, provides methods and techniques for measuring human abilities and assessing and improving the quality of these measurements. The process of measurement may be regarded as one of inferring an examinee's ability, which is not directly observable, from his observable responses to a test or a set of tests. For such an inference to be possible, test models must be developed that postulate certain statistical relationships between the underlying ability to be measured and the observed responses or test scores.

Two subdivisions of mental test theory—weak true-score theory and strong true-score theory—can be distinguished in terms of the strength of assumptions postulated in the test models in these theories. In weak true-score theory, which includes classical test theory and generalizability theory, the true score is defined as a certain type of expected value of the observed-score random variable, but no assumptions are made concerning the functional form of the conditional distribution of the observed score for a given true score. (Although some distributional assumptions are sometimes made in applications, they are

not essential to weak true-score theory.) Weak true-score theory is widely applicable to actual test data and has provided a large body of important results and techniques (Lord & Novick, 1968, chaps. 3-11; Novick, 1966).

Strong true-score theory is a collection of theories and models in which a certain functional form is assumed for the conditional distribution of the observed score for a given true score. This permits stronger inferences about the abilities of examinees and the properties of tests (or test items) than can be obtained from weak true-score theory models (Lord & Novick, 1968, chaps. 16-24). Item response theory, with which the present study is concerned, is a special case of strong true-score theory, where the conditional distributions of scores on individual items for a given true score or ability play a central role. Strong true-score theory also includes binomial error models (Lord, 1965b) and Rasch's (1960) Poisson process models.

Item response theory, also referred to as latent trait theory and item characteristic curve theory, was given formality and substance as a theory by Lord (1952) on the basis of earlier contributions by Ferguson (1942), Lawley (1943, 1944), Tucker (1946, 1948), and Lazarsfeld (1950), among others. Most of the theoretical developments and applications of item response theory have dealt with the cases where the test items are dichotomously scored and the ability to be measured is unidimensional. (For some important exceptions, see the references cited in Lord, 1980, chap. 2.) That is, it is usually (and also in the present paper) assumed that the items are scored 0 or 1 when the responses are wrong or

right, respectively, and that the ability of each examinee to be measured is represented by a single real number on the scale called the ability scale or the latent trait scale.

The fundamental assumption in item response theory is the assumption of local independence which requires that, at any fixed level of ability, all item responses be mutually statistically independent. As Lazarsfeld and Henry (1968, p. 22) pointed out, the local independence assumption is crucial not only from a statistical point of view but also from a substantive point of view. From the first viewpoint, this assumption provides mathematical simplicity in estimating examinees' abilities using the likelihood of item response patterns (Birnbaum, 1968, chap. 20), in deriving the conditional distribution of test scores at a given level of ability from the conditional distributions of item scores (Lord, 1952, sec. B1), and in other theoretical developments.

From the second viewpoint, the assumption implies that the heterogeneity among the examinees in a certain population on the ability under study is all accounted for by the derived ability scale, since the heterogeneity is considered to be responsible for statistical dependence among item responses in the population. Thus, the ability scale in item response theory is seen to be much like the factor score scale in the factor analytic model. However, the fundamental notion in the latter model is conditional linear independence rather than conditional statistical independence (Jöreskog, 1979).

Ability Scale in Item Response Theory

Provided that the assumption of local independence is satisfied for a certain pool of items, the unidimensional ability scale in item response theory is equivalent, up to monotone transformation, to the true score scale as it is defined in the classical test theory model for the test consisting of these items. The metric of the scale is usually determined so that the conditional probability of a correct answer to each item for a given ability level can be expressed as a simple mathematical function of the ability. The conditional probability of a correct answer, when expressed as a function of ability, is called the item response function, item characteristic function, or trace line.

The item response function just defined also represents the (curvilinear) regression of the dichotomous item score on the ability. Several item response models are distinguished from each other by the different mathematical forms which they assume for the item response functions. For example, the logistic item response model, on which the present study is focused, assumes that every item response function is given by a modified form of the logistic cumulative distribution function with three (item) parameters. (This model will be explained in more detail in Chapter II.)

Although the metric of the ability scale is determined once a mathematical form of item response functions is selected (e.g., logistic), the origin and unit of the scale are still arbitrary: Any change in the origin and unit of the scale is neutralized by appropriate linear transformations of item parameters, without altering the degree

of model fit or the general form of the item response functions (see Chapter II).

Thus, when the abilities of a sample of examinees and the parameters of a set of items are jointly estimated, or when the item parameters are estimated without knowledge of the abilities of the examinees in the sample, it becomes necessary to set the mean and variance of ability estimates in the sample or ability parameters in the population at some arbitrary values (e.g., 0 and 1, respectively). An important consequence of the arbitrariness of the origin and unit of the ability scale is that the ability estimates of two examinees obtained on two separately derived scales are not directly comparable, even if the two scales represent precisely the same underlying ability. Likewise, the parameters of two items defined on separately derived scales are not directly comparable.

Equating of Ability Scales

There are many instances in practice where it is necessary to transform an ability scale so that the transformed ability parameters become directly comparable with the parameters on another separately derived scale. In this paper, such a scale transformation process is referred to as equating of ability scales. (Although the term equating commonly refers to transformations performed on test scores rather than underlying abilities, it may be used more generally to stand for any type of scale transformation for equivalence.) Once the equating function which is used to transform one ability scale to another is

obtained, all item parameters on the first scale can be transformed accordingly so that the transformed parameters become directly comparable with those on the second scale [see (11) and (15) in Chapter II].

Equating of ability scales has been performed for various research purposes by many researchers, including Shiba (1978, construction of a broad-range vocabulary test), Laksana and Coffman (1980, item bias study), Loyd and Hoover (1980, vertical equating of test scores), Petersen, Cook, and Stocking (1981, horizontal equating of test scores), and Albanese (1981, study of robustness of item response models). The equating design used in the studies cited here is such that two or more different test forms (or levels) linked together by pairwise common items are administered to different samples of examinees, and ability scales are derived separately for different test forms. Then all scales are equated to one of the original scales which has been selected as a common scale, using the information obtained from the responses of two samples of examinees to each of the pairwise common items.

The basic equating design to be assumed in the paper is similar to those used in the studies cited above, but it is more general in that certain items may be common to more than two test forms. (Test forms, test levels, or more generally, sets of items to be administered will hereafter be referred to as test forms.) The common items may or may not be shared by all test forms: If they are common to all forms, they may be called the anchor items; if not, they may be called the locally common items, including the pairwise common items. In the subsequent

discussion, however, any of these types of items will be referred to simply as the common items when the type is understood from the context or when the distinction between the types is irrelevant.

Some of the equating designs belonging to the basic equating design will be introduced in Chapter II (see Figures 1 through 4) and Chapter IV (see Figures 5 through 7). Equating designs which do not belong to the basic design include the one in which ability scales are separately derived from non-overlapping test forms given to the same group or randomly equivalent groups of examinees. See Marco (1977) and Cook and Eigner (1981) for discussion on these designs and Kolen (1981) for a related application.

Purposes of the Study

This study has two major purposes. The first is to provide a detailed exposition of a general method for equating two or more ability scales which has been recently developed by the author. The method allows many different variations of the basic equating design described in the preceding section. The logistic item response model is assumed throughout the paper, although some of the general discussion is applicable to other models as well. In the next chapter, the model and the problem are formally presented, and several equating methods which have been previously developed and used are reviewed. Part B is devoted to the description of the new method: A special case of the method is treated in Chapter III, and a general case is discussed in Chapter IV.

The second major purpose of the study is to evaluate the method by means of computer simulation. Chapter V of Part C describes the procedures used for generating necessary data sets and evaluating the results. The results of the numerical evaluation are presented and discussed in Chapter VI.

Part A and Part B of the present paper overlap two previous articles (Haebara, 1980, 1981) in which the method was proposed for two specific applications, but the present paper provides a far more complete treatment of the method in particular and of equating in general. First of all, the method itself has been generalized so that the techniques proposed in the previous papers as two different methods can now be treated as two special (and simplest) cases of the general method presented in Chapter IV. In addition, an index of internal consistency of an equating system is introduced in connection with the general method. Furthermore, the notion of item-specific weight functions, which may be the key to the future improvement of the method, is introduced (Chapter III). The present paper also covers equating methods recently developed by Lord (1981) and Divgi (in press), which may be regarded as modifications of the method presented here.

CHAPTER II

FORMULATION OF THE PROBLEM

Logistic Item Response Model

Let θ denote the ability to be measured. Then, in the logistic item response model, the conditional probability of answering an item correctly for given θ is assumed to be represented by

$$P(\theta) = c + (1 - c)/\{1 + \exp[-Da(\theta - b)]\}, \qquad (1)$$

where a, b, and c are parameters characterizing the item, and D is a scaling constant which is commonly given the value of 1.7 (the reason for assigning this particular value will be explained later). The range of θ is $(-\infty, \infty)$ in (1) and throughout the paper. The item response function $P(\theta)$ is a monotonically increasing function of θ and has an inflection point at $\theta = b$. The parameter b is called the item difficulty and determines the relative position of the curve along the θ scale. The slope of the curve at the inflection point is proportional to the parameter a (when c is held constant), and this parameter is called the discriminating power of the item. Finally, the parameter c represents the height of the curve at $\theta = -\infty$ and is called the guessing parameter.

In applications of this model, the guessing parameter c is sometimes fixed a priori at zero or at a certain positive constant for all items. When the parameter c is fixed at zero for all items, this reduced model is called the two-parameter logistic model, since there are only two parameters left in the model to characterize different items. In addition to fixing c at zero, the discriminating power a is sometimes fixed a priori at a certain positive constant for all items. This further simplified model is called the one-parameter logistic model or the Rasch model, since the model becomes equivalent to one of the test models developed by Rasch (1960). The treatment of the logistic model in this paper will assume the general case as given in (1).

Historically, the general (three-parameter) logistic model of the form given in (1) was developed by Birnbaum (1968) as an extension of the two-parameter logistic model (Birnbaum, 1958) to deal with the situation in which it is possible to give correct item responses by guessing. The two-parameter logistic model, on the other hand, was originally developed as a mathematically tractable substitute for the (two-parameter) normal ogive model, which had been used by the pioneers of item response theory (Lord, 1952; Tucker, 1946; etc.). The generalization of the normal ogive model to deal with guessing, in which the three parameters a, b, and c are given the same interpretation as in the logistic model, was also described by Birnbaum (1968). For the same values of a, b, and c, the logistic function best approximates the normal ogive when the scaling constant D in (1) is about 1.7, in which case the absolute difference between the two functions is always less

than 0.01 (Birnbaum, 1968, p. 399). Despite the closeness of these two functions, however, there are also very important differences in their statistical properties (Birnbaum, 1968; Samejima, 1969, chap. 10).

Recently, Samejima (1979, 1980) has developed a new family of item response models for multiple-choice items, based on earlier developments on models for multi-category item response data by Samejima (1969, 1972) and Bock (1972). Her models provide operating characteristics of, i.e., conditional probabilities of responding to, individual alternatives in each item, and the operating characteristic of the correct alternative corresponds to the item response function. One of the new models, which is referred to as Model B of Type I (Samejima, 1979, p. 17), for example, assumes an item response function similar to the logistic function but allows non-monotony of the function. More specifically, in this and other models in the family, the asymptote of the item response function at the lower extreme of the ability scale is given by the reciprocal of the number of alternatives in the item, and the function first decreases up to a certain point on the θ scale and then increases toward the upper asymptote of 1.0. These models seem to be more realistic and to provide a better fit to such items as those reported as anomalies by Lord (1965a) than the logistic or the normal ogive model. However, further investigation still needs to be done for the usefulness of these models to be properly assessed.

When the logistic item response model is applied to practical measurement problems, the item parameters a, b, and c must be estimated for each item, and, in most applications, the ability parameter θ of

each examinee must also be estimated. Estimation procedures which have been proposed and used include the maximum likelihood method (Birnbaum, 1968; Lord, 1968; Wood, Wingersky, & Lord, 1976), the marginal maximum likelihood method (Bock, 1981; Bock & Lieberman, 1970), the Bayesian hierarchical method (Swaminathan & Gifford, 1981), and other methods (Jensema, 1976; Urry, 1974, 1978; see also McKinley & Reckase, 1981; Ree, 1979). However, since the new equating method to be presented and investigated in this paper is applicable no matter what estimation procedure may be used, no further discussion will be provided on the problem of item and ability parameter estimation. Nevertheless, the importance of reliable estimation of these parameters can never be overemphasized in any application of item response models.

Equating Problem and Some Remarks

In Chapter I, it was pointed out that the logistic ability scale has arbitrary origin and unit. The following derivation gives a formal explanation of this property of the scale. Consider first an ability scale θ and parameters a, b, and c of an item. The conditional probability, given θ, of answering this item correctly is assumed to be given by (1). Next, consider an arbitrary linear transformation of the θ scale defined in the form of

$$\theta^* = k\theta + d, \tag{2}$$

where θ^* is the transformed scale, and k and d are arbitrary constants such that $0 < k < \infty$ and $-\infty < d < \infty$. Consider also the following set of linear transformations of the item parameters:

$$\begin{cases} a^* = a/k \\ b^* = kb + d \\ c^* = c. \end{cases} \tag{3}$$

Then, using (2) and (3), Equation (1) can be expressed as

$$\begin{aligned} P(\theta) &= c + (1 - c)/\{1 + \exp[-Da(\theta - b)]\} \\ &= c + (1 - c)/[1 + \exp\{-D(a/k)[(k\theta + d) - (kb + d)]\}] \\ &= c^* + (1 - c^*)/\{1 + \exp[-Da^*(\theta^* - b^*)]\}. \end{aligned} \tag{4}$$

Thus, the item response function $P^*(\theta^*)$ for the transformed ability θ^* is given by

$$P^*(\theta^*) = c^* + (1 - c^*)/\{1 + \exp[-Da^*(\theta^* - b^*)]\}, \tag{5}$$

which is also a logistic item response function. Therefore, the ability scales θ in (1) and θ^* in (5) are both logistic ability scales, and yet they have different origins and units as implied by (2). This is the exact meaning of the arbitrariness of the origin and unit of the logistic ability scale.

Next, let us observe another property of the logistic ability scale, which is closely related to the one discussed in the preceding paragraph. This property can be stated as follows. Separately derived logistic ability scales measuring the same underlying ability but having different origins and units are different only in their origins and

units, and therefore, they can be equated by appropriate linear equating functions. This is proved now for the case of two scales. Suppose an item has its item parameters defined on two scales, and let [a, b, c, θ, P(θ)] and [a*, b*, c*, θ*, P*(θ*)] denote the sets of the parameters of the item, the ability parameter, and the item response function defined on the two scales. Assuming that the item parameters are true values (not estimates), we first note that an ability parameter θ on the first scale represents the same ability level as an ability parameter θ* on the second scale if and only if

$$P(\theta) = P^*(\theta^*). \tag{6}$$

That is, the equivalence between θ and θ* is determined in terms of the probabilities of a correct answer associated with these ability parameters. We also note that, for Equation (6) to hold for all ability levels, the probability of an examinee with an infinitely low ability level answering the item correctly must be the same for the two scales; i.e.,

$$c = P(\theta|\theta = -\infty) = P^*(\theta^*|\theta^* = -\infty) = c^*. \tag{7}$$

Then, from (6), (7), and (1), we obtain

$$a(\theta - b) = a^*(\theta^* - b^*). \tag{8}$$

Equation (8) is equivalent to

$$\theta^* = k\theta + d, \tag{9}$$

where

$$\begin{cases} k = a/a^* \\ d = b^* - (a/a^*)b. \end{cases} \tag{10}$$

Thus, the equating from θ to θ^* (and from θ^* to θ) must in fact be linear.

Suppose now that there are m (≥ 2) separately derived logistic ability scales θ_j ($j = 1, 2, \ldots, m$) measuring the same underlying ability. From the foregoing discussion, it is clear that all of these scales can be linearly equated to each other. Moreover, if one of the scales, say Scale 1, is chosen as the common scale to which all other scales are to be equated, we can set m linear equating functions of the form

$$\theta_1 = k_j\theta_j + d_j, \qquad j = 1, 2, \ldots, m, \tag{11}$$

where k_j and d_j are the equating constants. [In (11), k_1 and d_1 must be 1 and 0, respectively, for an obvious reason.] The equating problem to be solved can now be clearly defined as the problem of estimating the equating constants k_j and d_j in (11) for all j.

To see the generality of the problem just formulated, suppose that Scale j' ($j' \neq 1$) was chosen as the common scale instead of Scale 1. Then the equating functions would be given by

$$\theta_{j'} = k_j^{**}\theta_j + d_j^{**}, \qquad j = 1, 2, \ldots, m, \tag{12}$$

where, again, $k_{j'}^{**} = 1$ and $d_{j'}^{**} = 0$. But we have, from (11),

$$\theta_{j'} = (\theta_1 - d_{j'})/k_{j'}$$

$$= [(k_j\theta_j + d_j) - d_{j'}]/k_{j'}$$

$$= (k_j/k_{j'})\theta_j + (d_j - d_{j'})/k_{j'}, \tag{13}$$

so that

$$\begin{cases} k_j^{**} = k_j/k_{j'} \\ d_j^{**} = (d_j - d_{j'})/k_{j'}. \end{cases} \tag{14}$$

Therefore, once the constants k_j and d_j in (11) are estimated for all j, they can be substituted into (14) to provide the estimates of the necessary equating constants in (12). Thus, it causes no loss of generality to consider only those equating functions that are applied to equate all scales to Scale 1.

Linear transformations of item parameters defined on the m original scales, which are necessary for making them comparable with each other on the common scale, can be easily obtained, once the equating constants k_j and d_j are estimated. Using an argument parallel to the one used in the first paragraph of this section, it can be shown that the necessary item parameter transformations are of the form of Equation (3). Introducing subscripts to item parameters to be consistent with the subsequent notation system, Equation (3) can be rewritten as

$$\begin{cases} a_{g,j}^{*} = a_{g,j}/k_j \\ b_{g,j}^{*} = k_j b_{g,j} + d_j \\ c_{g,j}^{*} = c_{g,j}, \end{cases} \tag{15}$$

where $a_{g,j}$, $b_{g,j}$, and $c_{g,j}$ are the original parameters of item g defined on Scale j (j = 1, 2,..., m), and $a_{g,j}^{*}$, $b_{g,j}^{*}$, and $c_{g,j}^{*}$ are the transformed parameters which are now defined on the common scale. [Note that the third equation in (15) is trivial: The c parameter is defined on the probability scale, and therefore, it should not be affected by any transformation on the θ scale.]

Equating Designs

As stated in Chapter I, the basic equating design assumed throughout this paper is the one in which there are m ($\geq$ 2) test forms linked together by common (pairwise common, locally common, or anchor) items and they are administered to m different samples of examinees to yield m logistic ability scales. There are many variations in this basic design and four such designs are illustrated in Figures 1 through 4 for m = 4. (Some additional designs are introduced in Chapter IV.) In these figures, boxes represent test forms, and line segments connecting them represent sets of common items.

In the open chain design in Figure 1, all but two forms at the ends are linked to two adjacent forms, and no forms are linked to more than two forms, implying that the items common to a pair of forms are not shared by any other forms. (Even if certain items are common to three

or more forms, the design would be regarded as the open chain if such higher-order overlaps were ignored in actual equating. This is, of course, not a good procedure to recommend.) In the open chain design, it is clear that all we need is an equating method which can be used to equate two scales derived from two test forms having some items in common, since repeated applications of such a method will eventually equate all scales. All of the previously developed equating methods to be reviewed in the next section have dealt with the two-scale equating problem. These methods can, therefore, be readily applied to the m-form open chain design. The open chain design is also assumed in the special case of the new method; in Chapter III, this special case will be presented first for the case of two scales and then its extension to m-scale equating will be discussed. As we shall see later, the two-scale equating problem is also basic to equating designs other than the open chain design.

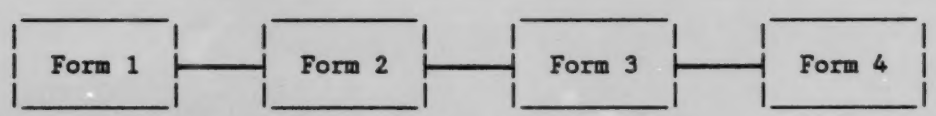

Figure 1: Open chain of four test forms linked by common items.

The closed chain design (Figure 2), the incomplete network design (Figure 3), and the complete network design (Figure 4) have more strongly interconnected linkage patterns than the open chain design. It must be intuitively clear that the stronger interconnections among test

forms will generally result in the more satisfactory equatings. Equatings that use these more desirable designs can be performed by the general method to be presented in Chapter IV. In that chapter, more discussion will be provided on these designs. Similar test form linkage patterns have been discussed in other contexts by Angoff (1971, p. 587), Lord (1980, pp. 154, 205), Petersen et al. (1981), Wilks (1961), and Wright (1977).

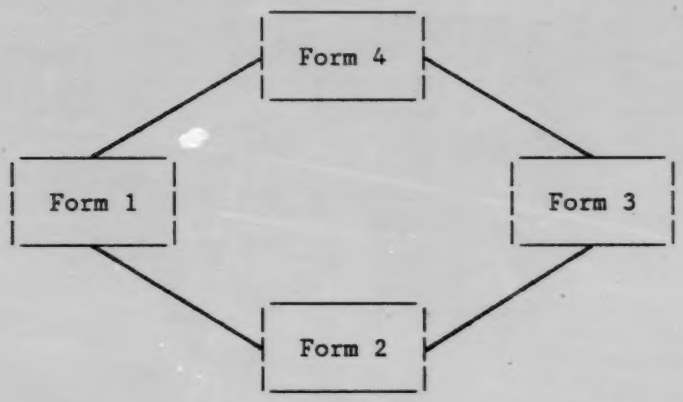

Figure 2: Closed chain of four test forms linked by common items.

Previously Developed Methods

In this section, previously developed methods for equating two logistic ability scales are reviewed. Let $P_{g,j}$ denote the item response function for common item g (g = 1, 2,..., r, say) defined on Scale j (j = 1, 2), and let $a_{g,j}$, $b_{g,j}$, and $c_{g,j}$ be the parameters of $P_{g,j}$. Then, following the same line of argument that led to Equations (6) through (10), we first note that a perfect equating should satisfy the relationship

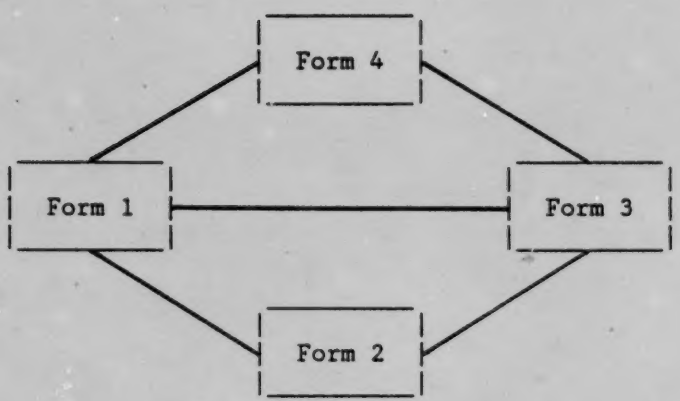

Figure 3: Incomplete network of four test forms linked by common items.

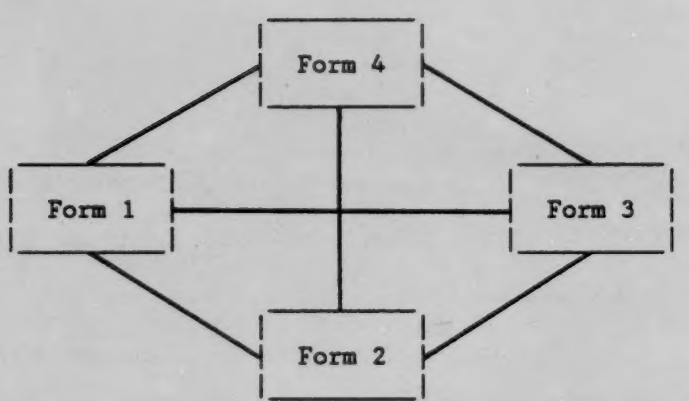

Figure 4: Complete network of four test forms linked by common items.

$$P_{g,1}(\theta_1) = P_{g,2}(\theta_2), \quad g = 1, 2, \ldots, r, \tag{16}$$

for $-\infty < \theta_1, \theta_2 < \infty$, where θ_1 and θ_2 are the equivalent (but numerically different) ability parameters on Scale 1 and Scale 2 that are equated by

$$\theta_1 = k\theta_2 + d. \tag{17}$$

[For simplicity, the subscript to k and d is omitted in (17) and other places where the two-scale equating problem is discussed.]

In practice, however, Equation (16) never holds exactly for all common items, no matter what values may be used for k and d in (17). This is because the item parameter estimates are subject to sampling fluctuations and because the model fit is inevitably less than perfect. For the same reasons, the values of k and d computed by

$$k = a_{g,2}/a_{g,1} \tag{18}$$

and

$$d = b_{g,1} - kb_{g,2}, \tag{19}$$

which are the solutions implied by (10), ordinarily vary, depending on which one of the common items is used in the computation. All the methods introduced here are aimed at estimating the values of k and d in such a way that Equation (18) and/or Equation (19) approximately holds for all g.

Marco (1977), for example, focused on (19) and suggested estimating k and d by

$$\begin{cases} \hat{k} = (\text{SD of } b_{g,1})/(\text{SD of } b_{g,2}) \\ \hat{d} = (\text{Mean of } b_{g,1}) - \hat{k}(\text{Mean of } b_{g,2}), \end{cases} \tag{20}$$

where the means and SD's (standard deviations) are taken over the r common items. Ree and Jensen (1980) conducted an empirical study regarding effects of sample size on these estimates of the equating constants. Douglass (1980) discussed the use of the coefficient of regression of $b_{g,1}$ on $b_{g,2}$. The estimate of k is then given by the product of $\hat{k}$ in (20) and the correlation between $b_{g,1}$ and $b_{g,2}$. The estimate of d is computed as in (20), using the obtained estimate of k.

Shiba (1978) focused on (18) and proposed estimation of k by using the scatter diagram of r points whose coordinates are $(a_{g,1}, a_{g,2})$. The estimate of k is then given by the slope of either the principal axis in the scatter diagram or the line that goes through the centroid of the points and the origin. Douglass (1980) suggested an alternative that uses the average of $a_{g,2}/a_{g,1}$ as the estimate of k. In either case, the estimate of d is computed as in (20), using the obtained estimate of k.

All these methods provide fairly simple ways of estimating the equating constants. A disadvantage common to them, however, is that a small number of poorly estimated item parameters may affect the equating result drastically. To overcome this difficulty, Linn, Levine,

Hastings, and Wardrop (1980, 1981) have proposed a modification of Equation (20), in which the means and SD's are replaced by the weighted means and weighted SD's. Their method uses, for each common item g, the reciprocal of the larger of the approximate sampling variances of the estimates of $b_{g,1}$ and $b_{g,2}$ as the weight given to $b_{g,1}$ and $b_{g,2}$. The approximate sampling variances of the item parameter estimates are computed by assuming that the estimates are obtained by the maximum likelihood method.

Petersen et al. (1981) have developed a similar but more sophisticated modification of (20). Their method employs an iterative procedure. First, initial weights for $b_{g,1}$ and $b_{g,2}$ (g = 1, 2,..., r) are computed as done by Linn et al. Second, the initial estimates of k and d are computed as in (20), using these initial weights. Third, the square of $[b_{g,1} - (\hat{k}b_{g,2} + d)]$ is computed for each g, and the product of a certain decreasing function of this quantity and the initial weight becomes the new weight for $b_{g,1}$ and $b_{g,2}$. Fourth, updated estimates of k and d are computed, using these new weights. Then the third and fourth steps are repeated until convergence is reached. This method gives low weights to those item difficulties that are poorly estimated or outliers, thus providing a more robust way of estimating the equating constants than the direct use of (20).

Although the last two methods are considered to be significant improvements over the earlier procedures, they still have an important drawback in that they are concerned only with the relationship (19) between item difficulties, completely ignoring the relationship (18)

between discriminating powers. More importantly, these and all other methods introduced in this section do not directly deal with the most fundamental relationship between item response functions which is given in (16). The best equating function must be a function that makes the equality in (16) hold as nearly as possible for all common items, since, as pointed out earlier, the equivalence between ability parameters on different scales is determined in terms of the conditional probabilities of correct answers associated with them. These considerations have led to the development of the new equating method to be presented in the next two chapters.

PART B

THE METHOD

CHAPTER III

UNRESTRICTED CASE

Least Squares Criterion

The new equating method is general enough to be applicable to various equating designs. In this chapter, however, only a special case of the general method intended specifically for the open chain design is presented. In this special case of the method, the equating constants are estimated in such a way that a certain criterion function is minimized with no restrictions being imposed on the estimates. Since, as mentioned earlier, equating in the open chain design basically requires only a procedure for equating one scale to another, just one equating function of the form

$$\theta_1 = k\theta_2 + d, \tag{21}$$

is considered here. (Extension to equating of more than two scales will be discussed later in this chapter.) Equating from Scale 1 to Scale 2 is, of course, given by the inverse transformation of (21); i.e.,

$$\theta_2 = (\theta_1 - d)/k. \tag{22}$$

To obtain such estimates of the equating constants k and d that provide the best approximation to the perfect equating (16), the errors of equating $e_{gi,1}$ and $e_{gi,2}$ are first defined by

$$e_{gi,1} = P_{g,1}(\theta_{i,1}) - P_{g,2}[(\theta_{i,1} - d)/k] \quad (23)$$

and

$$e_{gi,2} = P_{g,2}(\theta_{i,2}) - P_{g,1}(k\theta_{i,2} + d), \quad (24)$$

where

$e_{gi,1}$ = The error of equating from Scale 1 to Scale 2 for the g-th common item and the i-th comparison point (i.e., ability parameter value at which the error is computed) on Scale 1,

$e_{gi,2}$ = The error of equating from Scale 2 to Scale 1 for the g-th common item and the i-th comparison point on Scale 2,

$P_{g,1}$ = The item response function of the g-th common item defined on Scale 1,

$P_{g,2}$ = The item response function of the g-th common item defined on Scale 2,

$\theta_{i,1}$ = The i-th comparison point on Scale 1, and

$\theta_{i,2}$ = The i-th comparison point on Scale 2.

The errors of equating $e_{gi,1}$ and $e_{gi,2}$ are zero for all g and all i if and only if a perfect equating is obtained, and they deviate from zero either positively or negatively as the actual equating deviates from the perfect equating. (The reason for defining both $e_{gi,1}$ and $e_{gi,2}$ instead of just one of them will be explained shortly.)

Next, an appropriate loss function L that evaluates the loss associated with the magnitude of equating error is selected. Among the commonly used loss functions, the squared-error loss function

$$L(e) = e^2 \quad (25)$$

is used here. This loss function is computationally convenient and considered to be appropriate for the present problem because each squared quantity (i.e., error of equating) is bounded between 0 and 1; that is, the conceivable problem with the squared-error loss of putting an extremely large penalty on large errors can be avoided in the present case (see Novick & Jackson, 1974, p. 6).

Then the criterion function Q to be minimized is defined as

$$Q = \sum_{g=1}^{r} \{ \sum_{i=1}^{n_1} L(e_{gi,1}) h_1(\theta_{i,1}) + \sum_{i=1}^{n_2} L(e_{gi,2}) h_2(\theta_{i,2}) \}$$

$$= \sum_{g=1}^{r} \{ \sum_{i=1}^{n_1} \{P_{g,1}(\theta_{i,1}) - P_{g,2}[(\theta_{i,1} - d)/k]\}^2 h_1(\theta_{i,1})$$

$$+ \sum_{i=1}^{n_2} [P_{g,2}(\theta_{i,2}) - P_{g,1}(k\theta_{i,2} + d)]^2 h_2(\theta_{i,2}) \}, \qquad (26)$$

where r is the number of common items, n_1 and n_2 are the numbers of comparison points chosen on Scales 1 and 2, and $h_1(\theta_{i,1})$ and $h_2(\theta_{i,2})$ are certain weight functions. The weight functions are useful because the same amount of equating error is usually considered to be more serious at some points on the ability scale than at other points. For example, the researcher may wish to minimize the errors in the region where the distribution of abilities of a certain population or sample of examinees is concentrated, at the cost of somewhat larger errors elsewhere. More discussion will be provided about the weight functions later in this chapter.

If the criterion function were defined either by

$$Q_1 = \sum_{g=1}^{r} \sum_{i=1}^{n_1} L(e_{gi,1})h_1(\theta_{i,1}) \qquad (27)$$

or by

$$Q_2 = \sum_{g=1}^{r} \sum_{i=1}^{n_2} L(e_{gi,2})h_2(\theta_{i,2}), \qquad (28)$$

then the equating from Scale 1 to Scale 2 obtained by minimizing Q_1 and the equating from Scale 2 to Scale 1 obtained by minimizing Q_2 generally would not be the inverse transformation of each other. This is true even if $h_1(\theta) = h_2(\theta)$ for all θ and $\theta_{i,1} = \theta_{i,2}$ for all i, because the actual comparison points relative to the location of item response functions are generally different in the two cases. Thus, the use of either Q_1 or Q_2 will lead to a result which is dependent on the direction of equating. The main reason for defining the criterion function Q as the sum of Q_1 and Q_2 is that this difficulty can then be avoided. The result obtained by minimizing Q is equivalent no matter which one of the scales may be equated to the other scale. Such equating may be called symmetric equating.

The weighted least squares criterion Q defined in (26) can be expressed more compactly, with the aid of vector-matrix representation, as

$$Q = (\underline{f} - \hat{\underline{f}})'\underline{W}(\underline{f} - \hat{\underline{f}}), \qquad (29)$$

where $\underline{f}$ is an $r(n_1 + n_2) \times 1$ vector containing the first terms $[P_{1,1}(\theta_{1,1}), P_{1,1}(\theta_{2,1}),$ etc.] in all the squared differences in (26) arranged in certain order, $\hat{\underline{f}}$ is an $r(n_1 + n_2) \times 1$ vector containing the corresponding second terms $\{P_{1,2}[(\theta_{1,1} - d)/k], P_{1,2}[(\theta_{2,1} - d)/k],$ etc.$\}$ arranged in the same order, and $\underline{W}$ is an $r(n_1 + n_2) \times r(n_1 + n_2)$ diagonal matrix containing the values of the weight functions associated with the squared differences arranged in the same order. This representation of Q is convenient in the subsequent treatment.

Minimization of Criterion Function

Provided that the parameter estimates of all common items have been computed, and that the $n_1 + n_2$ comparison points and the weight functions have been specified by the researcher, the criterion function Q is a function of the equating constants k and d alone. Then, by the method of least squares, those values of k and d that minimize Q are taken to be the estimates of the equating constants; that is, the estimates $\hat{k}$ and $\hat{d}$ must be such that

$$Q(\hat{k}, \hat{d}) \leq Q(k, d), \tag{30}$$

where $\hat{k} \in \{k | 0 < k < \infty\}$ and $\hat{d} \in \{d | -\infty < d < \infty\}$.

Under a rather weak assumption that the global minimum of Q is at a stationary point, the relationship (30) leads to

$$\begin{cases} \partial Q/\partial \hat{k} = 0 \\ \partial Q/\partial \hat{d} = 0. \end{cases} \tag{31}$$

By defining a 2 x 1 vector $\underline{x}$ of estimates by

$$\underline{x} = \begin{bmatrix} \hat{k} \\ \hat{d} \end{bmatrix}, \tag{32}$$

the two equations in (31) can be expressed as

$$\partial Q/\partial \underline{x} = \underline{0}. \tag{33}$$

From (29), Equation (33) is equivalent to

$$\underline{g}(\underline{x}) \equiv -(1/2)(\partial Q/\partial \underline{x}) = (\partial \hat{\underline{f}}/\partial \underline{x})'\underline{W}(\underline{f} - \hat{\underline{f}}) = \underline{0}. \tag{34}$$

To solve (34) for $\underline{x}$, a modified form of the Gauss-Newton method (see Kennedy & Gentle, 1980, sec. 10.3.2) is applied which successively updates the vector of estimates by

$$\underline{x}^{(i+1)} = \underline{x}^{(i)} + t\underline{H}(\underline{x}^{(i)})^{-1}\underline{g}(\underline{x}^{(i)}), \tag{35}$$

where

$$\underline{H}(\underline{x}) = (\partial \hat{\underline{f}}/\partial \underline{x})'\underline{W}(\partial \hat{\underline{f}}/\partial \underline{x}), \tag{36}$$

and the parenthesized superscripts represent iteration numbers. The scalar t in (35) is a step-shortening factor which is set to one at the beginning of each iteration and halved whenever the criterion function Q increases with the current value of the factor. (In practice, however, halving is rarely necessary.) Any of the previously developed methods for estimating k and d (Chapter II) may be used to compute the initial values.

Every time the vector $\underline{x}$ is updated by (35), the first derivative of $\hat{\underline{f}}$ with respect to $\underline{x}$ must be evaluated. The $r(n_1 + n_2) \times 2$ matrix $\partial\hat{\underline{f}}/\partial\underline{x}$ has as its (i, j) element the derivative of the i-th element of $\hat{\underline{f}}$ with respect to the j-th element of $\underline{x}$. The following are the necessary formulas for these derivatives.

$$\begin{cases} \partial P[(\theta - d)/k]/\partial k = -[(\theta - d)/k^2]P'[(\theta - d)/k] \\ \partial P[(\theta - d)/k]/\partial d = -(1/k)P'[(\theta - d)/k]. \end{cases} \tag{37}$$

$$\begin{cases} \partial P(k\theta + d)/\partial k = \theta P'(k\theta + d) \\ \partial P(k\theta + d)/\partial d = P'(k\theta + d). \end{cases} \tag{38}$$

In (37) and (38), P' denotes the first derivative of the logistic item response function (1); that is, with a general argument z,

$$P'(z) = \partial P(z)/\partial z$$

$$= (1 - c)Da \cdot \exp[Da(z - b)]/\{1 + \exp[Da(z - b)]\}^2. \tag{39}$$

In the formulas above, subscripts and carets (^) are omitted to avoid complexity.

Remarks on Weight Functions

As pointed out earlier in this chapter, the weight functions in the least squares criterion serve the important purpose of specifying the relative seriousness of equating errors at different points on the ability scales. A reasonable way of specifying the weights is to use

the sample frequency distributions (grouped or ungrouped) of ability estimates in appropriate examinee samples. Alternatively, one of the standard probability density functions may be used either as an approximation to the sample frequencies or as representation of the researcher's subjective weights. If the researcher feels that equating errors are of equal importance at all points in a certain range on the ability scale, then the uniform distribution over that range may be used. In this case, of course, the result to be obtained is an unweighted least squares solution.

Although the present formulation of the criterion function allows just one weight function for each original scale, it is not difficult to make the least squares criterion allow different weight functions for different items. An advantage of using different weights for different items is that the reliabilities of the estimated item response functions can be taken into account; that is, those item response functions which are better estimated can be given more weight than others. If the item-specific weight functions that reflect the reliabilities of the estimated item response functions can be determined, then these weights may be multiplied by the scale-specific weights to yield the final weight functions.

The reliabilities of an item response function at various ability levels generally depend on the particular estimation procedure used. Linn et al. (1980) have proposed a formula for approximate standard errors of logistic item response functions estimated by the maximum likelihood method (see Wood et al., 1976). The formula was developed by

approximating an estimated item response function by a linear combination of the estimates of the three item parameters and by computing the variance of that linear composite under the assumption that the three item parameter estimates have a multivariate normal distribution with covariance matrix equal to the inverse of the information matrix. The necessary computations are fairly easy to perform.

As an example of the result obtained by the formula proposed by Linn et al., Table 1 shows the values of two hypothetical item response functions and the corresponding approximate standard errors computed at different ability levels. The two items are different only in their item difficulties ($b = -1.0$ for item 1 and $b = 1.0$ for Item 2). The common discriminating power and guessing parameter are 1.0 and 0.15, respectively. The computations were performed by assuming that there were only nine ability levels shown in the first column of the table, and that the relative frequencies of θ in the hypothetical sample of size 1000 were proportional to the ordinates of the standard normal probability density function at these points.

The table indicates that the standard errors are considerably different for the two items and also for different ability levels. For Item 1 (easy item), the standard error strictly decreases as a function of θ, from about 0.11 at $\theta = -4.0$ to about 0.00 at $\theta = 4.0$. On the other hand, the standard error for Item 2 (hard item) takes the largest value of about 0.03 at $\theta = 2.0$ and the smallest value of about 0.00 at $\theta = 4.0$.

Table 1

Approximate Standard Errors of Two Item Response Functions at Nine Ability Levels

	Item 1		Item 2	
	a = 1.0 b = -1.0 c = 0.15		a = 1.0 b = 1.0 c = 0.15	
θ	P(θ)	S.E.	P(θ)	S.E.
-4.0	0.1552	0.1077	0.1502	0.0302
-3.0	0.1775	0.0932	0.1510	0.0296
-2.0	0.2813	0.0525	0.1552	0.0271
-1.0	0.5750	0.0291	0.1775	0.0197
0.0	0.8687	0.0135	0.2813	0.0199
1.0	0.9726	0.0071	0.5750	0.0268
2.0	0.9949	0.0024	0.8687	0.0323
3.0	0.9991	0.0006	0.9726	0.0141
4.0	0.9998	0.0002	0.9949	0.0040

Taking these differential standard errors of estimated item response functions into account should lead to a more satisfactory solution. However, adequacy of the formula developed by Linn et al. (1980) has not yet been fully investigated. Moreover, the formulas applicable to parameter estimation procedures other than the maximum likelihood method are not yet available. Considering the insufficiency of these basic developments necessary for profitable use of item-specific weights, the remainder of the present paper will deal only with scale-specific weight functions. But it must be emphasized that this does not imply denial of potential advantages of using item-specific weight functions.

Related Methods

The least squares method being discussed is characterized by its explicit definition of a criterion function and minimization of that function. The estimation process having such characteristics is usually referred to as the optimization process. Following the development of the present method, two related methods have been proposed by Lord (1981) and Divgi (in press) which also have the characteristics of the optimization process. In the notation used in this paper, both of these methods define the error of equating e_i, for equating Scale 2 to Scale 1, by

$$e_i = \sum_{g=1}^{r} e_{gi,2}$$

$$= \sum_{g=1}^{r} P_{g,2}(\theta_{i,2}) - \sum_{g=1}^{r} P_{g,1}(k\theta_{i,2} + d)$$

$$= T_2(\theta_{i,2}) - T_1(k\theta_{i,2} + d), \tag{40}$$

where $T(\theta)$ is the sum of the item response functions over all common items. [If the set of common items is regarded as a test, $T(\theta)$ is called the test characteristic function.]

In Lord's (1981) method, a least squares criterion function, say Q_L, is then defined by taking the sum of the squares of e_i over a sample of examinees; that is,

$$Q_L = \sum_{i=1}^{n_2} e_i^2, \tag{41}$$

where n_2 is the number of examinees who took Form 2. Then the estimates of k and d are obtained by minimizing the criterion function. To relate this function to the one given in this paper, we note that

$$Q_L = \sum_{i=1}^{n_2} \left(\sum_{g=1}^{r} e_{gi,2} \right)^2$$

$$= \sum_{i=1}^{n_2} \sum_{g=1}^{r} e_{gi,2}^2 + \sum_{i=1}^{n_2} \sum_{g=1}^{r} \sum_{\substack{g'=1 \\ g \neq g'}}^{r} e_{gi,2} e_{g'i,2}. \tag{42}$$

The first term in (42) is now recognized as identical to Q_2 in (28) where $h_2(\theta) = 1$ for all θ. Thus, denoting the second term by Q_{L2}, we have

$$Q_L = Q_2 + Q_{L2}. \tag{43}$$

It is now clear that Lord's criterion function is different from Q_2 by an additional term which is the sum of the cross products of $e_{gi,2}$. To satisfy the requirement of symmetric equating mentioned earlier, Lord suggested equating in two directions (i.e., Scale 2 to Scale 1 and Scale 1 to Scale 2) and taking the bisector of the two equating functions.

Unlike Lord's method, Divgi's (in press) method adopts a minimax criterion so that the equating constants can be estimated independently of sample frequencies of ability estimates or subjectively chosen ability distributions (i.e., weight functions). Specifically, his method estimates k and d so that the maximum value of equating error in (40) over the entire range of the ability scale is minimized for these estimates. This method provides a symmetric equating.

An important difference between the present method and these two methods, which, in the author's opinion, illustrates an advantage of the present method, is noted below. Suppose, for certain values of k and d, that the error of equating $e_{gi,2}$ in (28) and (40) is zero for r - 2 common items and for all i. Also suppose, for the remaining two items (say, Item x and Item y), that the following relationships hold:

$$\begin{cases} P_{x,2}(\theta_{i,2}) = P_{y,1}(k\theta_{i,2} + d) \\ P_{y,2}(\theta_{i,2}) = P_{x,1}(k\theta_{i,2} + d) \end{cases} \tag{44}$$

for all i, and

$$\begin{cases} P_{x,2}(\theta_{i,2}) \neq P_{x,1}(k\theta_{i,2} + d) \\ P_{y,2}(\theta_{i,2}) \neq P_{y,1}(k\theta_{i,2} + d) \end{cases} \tag{45}$$

for at least some i. Then the value of Q_2 is not zero, reflecting the relationship in (45), and therefore, the criterion function Q is positive. On the contrary, the value of e_i in (40) becomes zero in this case, because, from (44), we have

$$P_{x,2}(\theta_{i,2}) + P_{y,2}(\theta_{i,2}) = P_{x,1}(k\theta_{i,2} + d) + P_{y,1}(k\theta_{i,2} + d), \tag{46}$$

despite the deviation from a perfect equating. Thus, the zero value of Lord's or Divgi's criterion function, unlike Q, does not necessarily mean the perfect equating.

Equating More Than Two Scales

As discussed in Chapter II, m - 1 scales may be equated to Scale 1 by (11). To estimate k_j and d_j for j = 2, 3,..., m by the least squares method for the open chain design, m - 1 equating functions which are used to equate the m - 1 pairs of adjacent scales are first defined:

$$\theta_{j-1} = k_j^*\theta_j + d_j^*, \qquad j = 2, 3, \ldots, m. \tag{47}$$

Then the method for equating two scales described in this chapter is applied to each of the equating functions in (47) to estimate k_j^* and d_j^* for $j = 2, 3, \ldots, m$. Finally, the estimates of k_j and d_j in (11) are computed by

$$\begin{cases} \hat{k}_j = \prod_{j'=2}^{j} \hat{k}_{j'}^* \\ \\ \hat{d}_j = \sum_{j'=2}^{j} (\hat{d}_{j'}^* \prod_{j''=2}^{j'-1} \hat{k}_{j''}^*), \end{cases} \tag{48}$$

where $\prod_{j''=2}^{j'-1} \hat{k}_{j''}^* = 1$ for $j' = 2$. These equations can be readily verified by successively combining the $m - 1$ equations in (47).

Common Item Response Functions

To place all original item parameters on the common scale (Scale 1), the estimates $\hat{k}_j$ and $\hat{d}_j$ are substituted into (15), where $\hat{k}_1 = 1$ and $\hat{d}_1 = 0$. However, final adjustments are necessary for common items, since each of the common items originally defined on Scale $j - 1$ and Scale j has two sets of item parameters computed by $(\hat{k}_{j-1}, \hat{d}_{j-1})$ and $(\hat{k}_j, \hat{d}_j)$. To obtain a common item response function for each of these items, a weighted least squares criterion function $Q_{g,j}$ for item g whose parameters are originally defined on Scale $j - 1$ and Scale j is defined by

$$Q_{g,j} = \sum_{i=1}^{n_{j-1}} [P_{g,j-1}^{*}(\theta_{i,j-1}^{*}) - P_{g,j}^{**}(\theta_{i,j-1}^{*})]^2 h_{j-1}^{*}(\theta_{i,j-1}^{*}) + \sum_{i=1}^{n_j} [P_{g,j}^{*}(\theta_{i,j}^{*}) - P_{g,j}^{**}(\theta_{i,j}^{*})]^2 h_{j}^{*}(\theta_{i,j}^{*}), \quad (49)$$

where $P_{g,j}^{**}$ is the common item response function to be estimated, $P_{g,j-1}^{*}$ and $P_{g,j}^{*}$ are computed from the item parameters obtained by (15), $\theta_{i,j-1}^{*}$ and $\theta_{i,j}^{*}$ are computed from $\theta_{i,j-1}$ and $\theta_{i,j}$ by

$$\begin{cases} \theta_{i,j-1}^{*} = \hat{k}_{j-1}\theta_{i,j-1} + \hat{d}_{j-1} \\ \theta_{i,j}^{*} = \hat{k}_{j}\theta_{i,j} + \hat{d}_{j}, \end{cases} \quad (50)$$

and finally h_{j-1}^{*} and h_{j}^{*} are given from h_{j-1} and h_{j} by changing the variables by (50). In words, $Q_{g,j}$ is the sum of the weighted squared distances from the two transformed item response functions to the common item response function to be estimated. The parameters $a_{g,j}^{**}$, $b_{g,j}^{**}$, and $c_{g,j}^{**}$ of $P_{g,j}^{**}$ are then estimated by minimizing $Q_{g,j}$ with respect to these parameters.

The modified Gauss-Newton method used earlier can be applied here in the minimization of $Q_{g,j}$. The first derivatives in the $(n_{j-1} + n_j) \times 3$ matrix $\partial\underline{\hat{f}}/\partial\underline{x}$ necessary to implement the algorithm are listed below. In the list, subscripts and carets are omitted again for simplicity, and P** = P**(θ*) and Q** = 1 - P**.

$$\begin{cases} \partial P^{**}/\partial a^{**} = D(\theta^{*} - b^{**})Q^{**}(P^{**} - c^{**})/(1 - c^{**}) \\ \partial P^{**}/\partial b^{**} = -Da^{**}Q^{**}(P^{**} - c^{**})/(1 - c^{**}) \\ \partial P^{**}/\partial c^{**} = Q^{**}/(1 - c^{**}). \end{cases} \tag{51}$$

The initial estimate for each of the parameters may be given by the average of the two transformed parameter values; i.e., the average of $a_{g,j-1}^{*}$ and $a_{g,j}^{*}$, etc. It must be noted here that the guessing parameter $c_{g,j}^{**}$ is often poorly estimated unless a number of comparison points are set at very low ability levels or large weights are put on these points. Thus, it may be advisable to fix $c_{g,j}^{**}$ at the initial value and to estimate the parameters $a_{g,j}^{**}$ and $b_{g,j}^{**}$ by minimizing $Q_{g,j}$ with respect to these two parameters only.

CHAPTER IV

GENERALIZATION OF THE METHOD

Another Approach to the Unrestricted Case

In Chapter III, equating of m ($\geq$ 2) scales in the open chain design was described as the direct extension of the two-scale equating. That is, the equating constants were first estimated separately for each pair of adjacent scales, and then these estimates were combined by (48) to yield the final equating constants for use in equating each scale to Scale 1. In this section, an alternative approach to the m-scale equating problem in the open chain design is provided which gives the same solution as before but which will lead to a generalization of the method.

Instead of first defining equatings between pairs of adjacent scales separately, we now start with the equating functions used to equate m - 1 scales to Scale 1; i.e.,

$$\theta_1 = k_j\theta_j + d_j, \qquad j = 2, 3, \ldots, m, \tag{52}$$

which we call the basic equating functions. Next, we attempt to express the equating functions for all pairs of adjacent scales in terms of the basic equating constants (i.e., the equating constants of the basic equating functions). This can be done by hypothesizing that Scale j is

first equated to Scale 1 by (52) and then equated back to Scale j - 1 by the inverse transformation of (52); that is,

$$\begin{aligned}\theta_{j-1} &= (\theta_1 - d_{j-1})/k_{j-1} \\ &= [(k_j\theta_j + d_j) - d_{j-1}]/k_{j-1} \\ &= (k_j/k_{j-1})\theta_j + (d_j - d_{j-1})/k_{j-1},\end{aligned} \tag{53}$$

for $j = 2, 3, \ldots, m$. When $j = 2$, we have $k_{j-1} = k_1 = 1$ and $d_{j-1} = d_1 = 0$ so that the adjacent-scale equating function is identical to the basic equating function in this case. Equating from Scale j - 1 to Scale j is given by

$$\begin{aligned}\theta_j &= [\theta_{j-1} - (d_j - d_{j-1})/k_{j-1}]/(k_j/k_{j-1}) \\ &= (k_{j-1}/k_j)\theta_{j-1} + (d_{j-1} - d_j)/k_j.\end{aligned} \tag{54}$$

Now that the adjacent-scale equating functions are defined by (53) and (54), they can be substituted into the formula for the criterion function given in (26). Specifically, the criterion function, say $Q_{(j-1)j}$, for Scales j - 1 and j is given by

$$\begin{aligned}Q_{(j-1)j} = \sum_{g=1}^{r_{(j-1)j}} \{ &\sum_{i=1}^{n_{j-1}} L(e_{gi,j-1})h_{j-1}(\theta_{i,j-1}) \\ &+ \sum_{i=1}^{n_j} L(e_{gi,j})h_j(\theta_{i,j})\},\end{aligned} \tag{55}$$

where

$$\begin{cases} L(e_{gi,j-1}) = \{P_{g,j-1}(\theta_{i,j-1}) \\ \qquad - P_{g,j}[(k_{j-1}/k_j)\theta_{i,j-1} + (d_{j-1} - d_j)/k_j]\}^2 \\ L(e_{gi,j}) = \{P_{g,j}(\theta_{i,j}) \\ \qquad - P_{g,j-1}[(k_j/k_{j-1})\theta_{i,j} + (d_j - d_{j-1})/k_{j-1}]\}^2, \end{cases} \tag{56}$$

and $r_{(j-1)j}$ is the number of common items shared by Forms j - 1 and j. This criterion function may be called the adjacent-scale criterion function. Then an overall criterion function T_u is defined by

$$T_u = \sum_{j=2}^{m} Q_{(j-1)j}, \tag{57}$$

where the subscript u is used to distinguish the unrestricted case from the general case to be introduced in the next section. Finally, the basic equating constants k_j and d_j (j = 2, 3,..., m) are estimated by minimizing T_u with respect to them.

In this approach, all the basic equating constants can be directly estimated without further adjustments such as those in (48). The equivalence between the results obtained by this approach and by the previous approach can be shown as follows. First, it is noted that the equating constants directly estimated in the two approaches are related by (48) or, equivalently, by

$$\begin{cases} k_j^* = k_j/k_{j-1} \\ d_j^* = (d_j - d_{j-1})/k_{j-1} \end{cases} \tag{58}$$

[compare (47) and (53)]. Let this one-to-one correspondence between the sets $A = \{k_j, d_j | j = 2, 3, \ldots, m\}$ and $B = \{k_j^*, d_j^* | j = 2, 3, \ldots, m\}$ be denoted by $B = f(A)$. Then, noting that the criterion function T_u takes the same value as the sum of the $m - 1$ criterion functions used in Chapter III for the corresponding values of A and B, it can now be concluded that, if a certain value of A, say A_0, minimizes the first quantity (i.e., T_u), then $B_0 = f(A_0)$ must minimize the second. The converse is also true.

The present approach to the m-scale equating problem in the open chain design has a great generality and can be applied to various equating designs. The next section discusses the generalization of the method and introduces the notion of internal consistency of an equating system, which is a crucial concept in equating designs other than the open chain design.

General Method

For a set of m test forms, there are many possible linkage patterns. The open chain design, on which we have focused up to this point, represents the simplest pattern, linking each test form to two adjacent forms except for the two forms at the ends of the chain. Thus, there are $m - 1$ links in the m-form open chain design. At the other extreme, the complete network design (see Figure 4 in Chapter II), in which all possible pairs of test forms are linked to each other by some common items, has $m(m - 1)/2$ links. Any other design with the number of links between $m - 1$ and $m(m - 1)/2$ may be called the incomplete network

design (Figure 3), of which the closed chain design (Figure 2) is a special case.

As examples of the incomplete design, let us consider two widely used multilevel tests, the Iowa Tests of Basic Skills (ITBS), Forms 7/8, and the Cognitive Abilities Test (CAT), Form 3. Table 2 shows how a total of 107 items of the ITBS, Test V (Vocabulary), are allocated to the 6 levels (Hieronymus, Lindquist, & Hoover, 1978, p. 14). Likewise, Table 3 shows the allocation of a total of 65 items of the CAT, Nonverbal Battery, Test 3, to the 8 levels (Thorndike & Hagen, 1979, p. 40). Figures 5 and 6 illustrate the equating designs (i.e., overlapping patterns) of these two subtests, respectively. While each design constitutes an incomplete network, the latter one has as many as 25 links and shows an extremely strong interlock of the levels (the complete network design would have 28 links in this case). In fact, if we consider only Levels A through F in Figure 6, then these 6 levels are seen to be constituting a complete network.

The general method for equating m scales to be outlined in this section is applicable to any of these different designs. (At this point, an explicit definition must be given to a term used in the subsequent discussion in order to avoid ambiguities and misunderstanding that may arise as a result of introducing equating designs different from the familiar open chain design. In the open chain design, each pair of adjacent test forms, i.e., Form j - 1 and Form j, are linked to each other so that the adjacent scales derived from them can be directly equated. In other designs, however, some of the adjacent forms may not

Table 2

Organization of the ITBS, Test V

Levels	Items
9	1 - 30
10	11 - 46
11	25 - 63
12	39 - 80
13	56 - 98
14	64 -107

Table 3

Organization of the CAT,
Nonverbal Battery, Test 3

Levels	Items
A	1 - 30
B	6 - 35
C	11 - 40
D	16 - 45
E	21 - 50
F	26 - 55
G	31 - 60
H	36 - 65

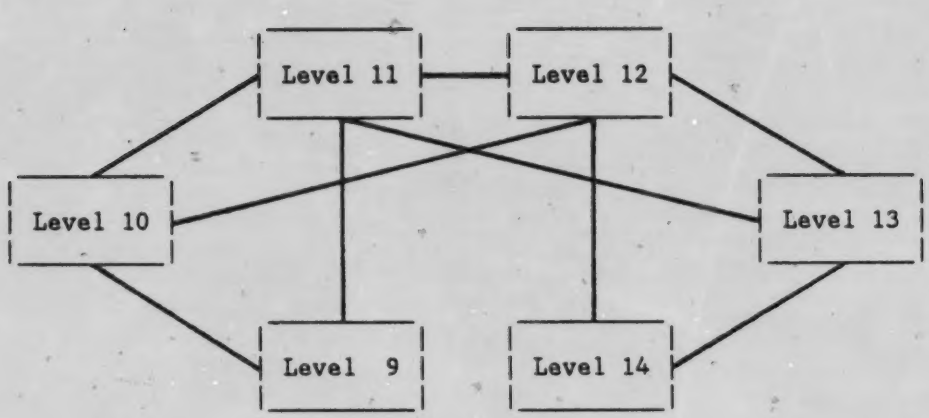

Figure 5: Incomplete network of 6 levels of the ITBS, Test V.

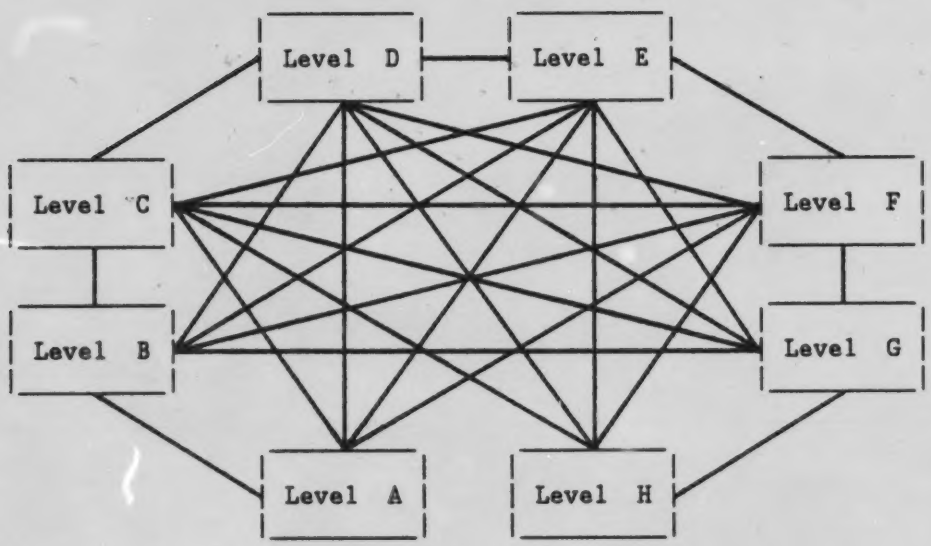

Figure 6: Incomplete network of 8 levels of the CAT, Nonverbal Battery, Test 3.

be linked by common items, while some of the non-adjacent forms may be linked. Hereafter, the term connected scales, instead of adjacent scales, will be used to refer to the scales that are derived from the linked forms and that can, therefore, be directly equated.)

When there are more than m - 1 links in the design, at least some scales can be equated to each other in at least two different ways. To make the discussion concrete, let us consider the four-form incomplete network design in Figure 3 in Chapter II. In this design, Scale 3 derived from Form 3, for example, can be equated directly to Scale 1, using the items common to Forms 1 and 3. However, Scale 3 can also be equated indirectly to Scale 1 through either Scale 2 or Scale 4. Thus, equating from Scale 3 to Scale 1 can be defined in three different ways. Suppose now that the equating functions for all pairs of connected scales are defined by

$$\begin{cases} \theta_1 = k_{12}\theta_2 + d_{12} \\ \theta_1 = k_{13}\theta_3 + d_{13} \\ \theta_1 = k_{14}\theta_4 + d_{14} \\ \theta_2 = k_{23}\theta_3 + d_{23} \\ \theta_3 = k_{34}\theta_4 + d_{34}. \end{cases} \tag{59}$$

Then the three equating functions for equating Scale 3 to Scale 1 are given by

$$\begin{cases} \theta_1 = k_{13}\theta_3 + d_{13} \\ \theta_1 = k_{12}(k_{23}\theta_3 + d_{23}) + d_{12} \\ \theta_1 = k_{14}[(\theta_3 - d_{34})/k_{34}] + d_{14}, \end{cases} \tag{60}$$

where the first equation represents the direct equating, and the second and the third represent the indirect equating through Scale 2 and Scale 4, respectively. It is necessary, of course, that these three equations define exactly the same scale transformation. This requirement may be called the internal consistency requirement. However, if the connected-scale equating constants in (59) were separately estimated by the least squares method or any other method, this crucial requirement generally would not be satisfied.

To satisfy the internal consistency requirement, therefore, it is clear that certain restrictions must be imposed upon the estimation of the equating constants. (The problem of internal consistency and the necessity for restricted estimation do not arise in the open chain design, where there are only m - 1 links. It is for this reason that the equating problem in the open chain design has been referred to as the unrestricted case. Two other examples of the unrestrictive design with m - 1 links are illustrated in Figure 7.) By comparing the coefficients in the three equations in (60), the necessary restrictions can be expressed as

$$\begin{cases} k_{13} = k_{12}k_{23} = k_{14}/k_{34} \\ d_{13} = k_{12}d_{23} + d_{12} = -(k_{14}/k_{34})d_{34} + d_{14}. \end{cases} \tag{61}$$

Restrictions of this type, however, may become very complicated as the number of forms and/or links increases.

Fortunately, however, there exists a much simpler way of estimating equating functions under the restriction for internal consistency. The restricted solution can be obtained by defining the equating functions for all pairs of connected scales not in terms of the pair-specific equating constants such as k_{12}, k_{23}, etc. in (59) but in terms of the basic equating constants as in (53) and (54). That is, as in the preceding section, the m - 1 basic equating functions are defined by (52) and then each equating between connected scales is expressed in terms of the constants in (52). Specifically, equating from Scale j to Scale j' is defined by

$$\theta_{j'} = (k_j/k_{j'})\theta_j + (d_j - d_{j'})/k_{j'}, \tag{62}$$

which has the inverse transformation given by

$$\theta_j = (k_{j'}/k_j)\theta_{j'} + (d_{j'} - d_j)/k_j. \tag{63}$$

Unlike the subscript j - 1 in (53) and (54), the subscript j' in (62) and (63) refers to any scale (adjacent or non-adjacent) that is connected to Scale j.

By noting that $k_1 = 1$ and $d_1 = 0$, it can be shown, from (62) or (63), that the first three equations in (59) are replaced by the basic equating functions

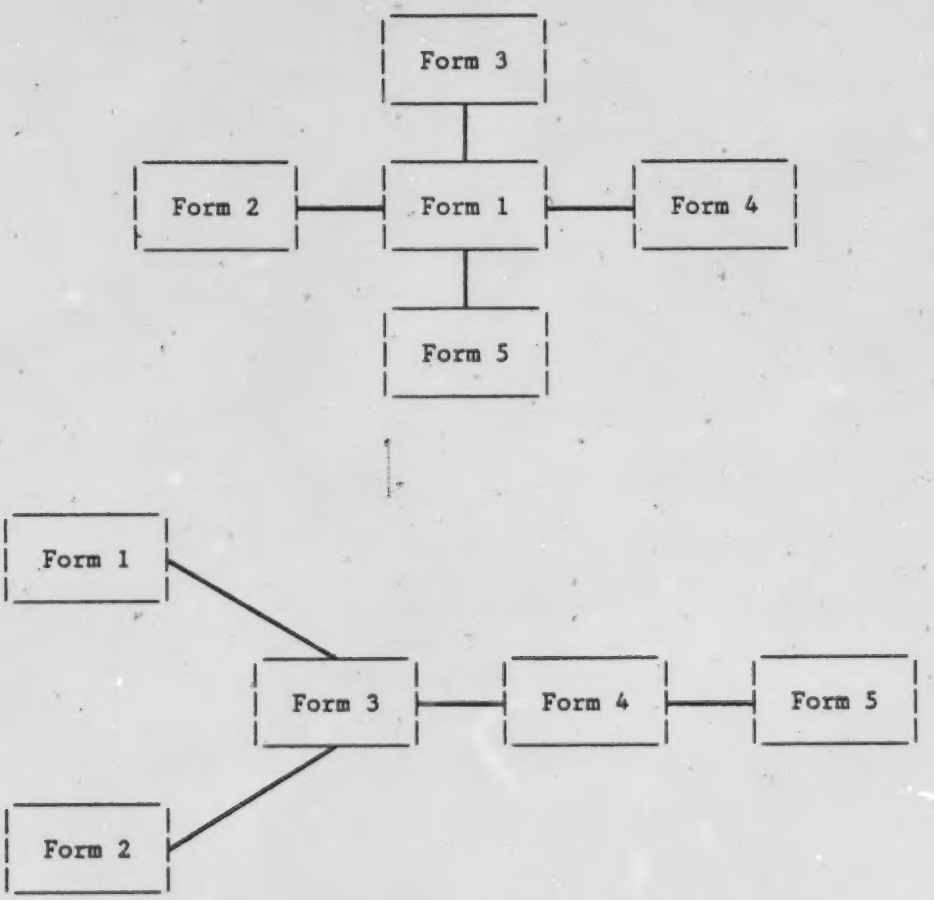

Figure 7: Two examples of unrestrictive design.

$$\begin{cases} \theta_1 = k_2\theta_2 + d_2 \\ \theta_1 = k_3\theta_3 + d_3 \\ \theta_1 = k_4\theta_4 + d_4. \end{cases} \tag{64}$$

The last two equations in (59) are replaced by

$$\begin{cases} \theta_2 = (k_3/k_2)\theta_3 + (d_3 - d_2)/k_2 \\ \theta_3 = (k_4/k_3)\theta_4 + (d_4 - d_3)/k_3. \end{cases} \tag{65}$$

From (64) and (65), the three equating functions for equating Scale 3 to Scale 1, which correspond to (60), are now given by

$$\begin{cases} \theta_1 = k_3\theta_3 + d_3 \\ \theta_1 = k_2[(k_3/k_2)\theta_3 + (d_3 - d_2)/k_2] + d_2 \\ \quad = k_3\theta_3 + d_3 \\ \theta_1 = k_4\{[\theta_3 - (d_4 - d_3)/k_3]/(k_4/k_3)\} + d_4 \\ \quad = k_3\theta_3 + d_3. \end{cases} \tag{66}$$

Thus, the internal consistency requirement is satisfied in this case. The requirement can be similarly shown to be satisfied in any arbitrary linkage pattern among m forms for any arbitrary set of basic equating constant values.

The next step in the general equating method is to define the overall criterion function T by

$$T = \sum_{j',j}^{\Sigma^*} Q_{j'j}, \tag{67}$$

where $Q_{j'j}$ is the connected-scale criterion function computed by (55) and (56) with replacement of the subscript $j - 1$ by j', and Σ^* stands for the summation over all pairs of connected scales. Then the final step is to jointly estimate all the basic equating constants k_j and d_j ($j = 2, 3, \ldots, m$) by minimizing T with respect to them.

Algorithm

The minimization of the overall criterion function T defined in (67) can be carried out by, again, applying the modified Gauss-Newton method used in Chapter III. In the present case, Equation (33) is replaced by

$$\partial T/\partial \underline{x} = \underline{0}, \tag{68}$$

where $\underline{x}$ is a $2(m - 1) \times 1$ vector of the estimates of the basic equating constants. To solve (68) for $\underline{x}$, it is successively updated by (35). The explanation of the algorithm given in Chapter III generally applies to the present case, except that the changes in the number of parameters to be estimated and in the number of squared differences alter the size of the vectors and matrices involved.

In particular, the matrix $\partial \hat{\underline{f}}/\partial \underline{x}$ is now a $\Sigma^* r_{j'j}(n_{j'} + n_j) \times 2(m - 1)$ matrix containing the first derivatives of $P_{g,j}[(\hat{k}_{j'}/\hat{k}_j)\theta_{i,j'} + (\hat{d}_{j'} - \hat{d}_j)/\hat{k}_j]$ and

$P_{g,j'}[(\hat{k}_j/\hat{k}_{j'})\theta_{i,j} + (\hat{d}_j - \hat{d}_{j'})/\hat{k}_{j'}]$ with respect to $\hat{k}_j$, $\hat{d}_j$, $\hat{k}_{j'}$, and $\hat{d}_{j'}$. The first derivatives of the first quantity are given below, with omission of certain subscripts, carets, and arguments.

$$\begin{cases} \partial P[(k_{j'}/k_j)\theta + (d_{j'} - d_j)/k_j]/\partial k_j \\ \qquad = -[(k_{j'}\theta + d_{j'} - d_j)/k_j^2]P'[(k_{j'}/k_j)\theta + (d_{j'} - d_j)/k_j] \\ \partial P/\partial d_j = -(1/k_j)P' \\ \partial P/\partial k_{j'} = (\theta/k_j)P' \\ \partial P/\partial d_{j'} = (1/k_j)P'. \end{cases} \tag{69}$$

In these formulas, the function P' is the first derivative of P given in (39), and $k_1 = 1$ and $d_1 = 0$ as before. The derivatives of the second quantity can be computed by interchanging j' and j in (69).

To place all original item parameters on the common scale and estimate a common item response function for each common item, the procedure described in the last section of Chapter III can be used. When certain items are common to more than two test forms, however, the criterion function given in (49) must be modified. Specifically, the criterion function in this case must be defined as the sum of the weighted squared distances from the three or more, instead of just two, transformed item response functions to the common item response function to be estimated.

Index of Internal Consistency

Equating functions for pairs of connected scales expressed in terms of the pair-specific equating constants were given in (59). In general, such a function for the pair of Scale j' and Scale j may be expressed as

$$\theta_{j'} = k_{j'j}\theta_j + d_{j'j}. \tag{70}$$

Then a criterion function $Q_{j'j}^*$ may be defined for this pair in the same way as $Q_{j'j}$ was defined, and an overall criterion function T* can be defined by taking the sum of $Q_{j'j}^*$ over all pairs of connected scales. Suppose now that T* is minimized with respect to the pair-specific constants $k_{j'j}$ and $d_{j'j}$. Then the minimized value of T* is always smaller than or equal to the minimized value of T in the preceding section computed from the same data; that is,

$$0 \leq \min(T^*) \leq \min(T). \tag{71}$$

This relationship follows from the general principle of the least squares estimation that a restricted solution can never provide a better fit, in terms of the minimized value of the criterion, than an unrestricted solution for the same data.

Let us now define an index ρ by

$$\rho = \min(T^*)/\min(T). \tag{72}$$

Then, from (71), we have

$$0 \leq \rho \leq 1. \tag{73}$$

If the unrestricted solution obtained by minimizing T* were identical to the restricted solution obtained by minimizing T, then this index would be unity. On the other hand, to the extent that these two solutions are different from each other, the value of ρ will be lower than unity. Since a large (negative) deviation of ρ from unity implies that the unrestricted solution is inconsistent, the index ρ may be called the index of internal consistency of the equating system. This index has a certain resemblance to the likelihood ratio (e.g., Hogg & Craig, 1978, sec. 7.4) used to test the tenability of a restrictive model, but the latter index is based on the maximum likelihood estimation instead of the least squares estimation.

As we have noted earlier, no restrictions can be imposed on the estimation in the open chain design or other designs with only m - 1 links (see Figure 7). Therefore, in this special case, we have

$$\min(T^*) = \min(T_u) = \min(T). \qquad (74)$$

and $\rho = 1$. However, this fact merely illustrates that it is impossible to check, by this index, the internal consistency in these designs. In other words, these designs have no degrees of freedom for inconsistencies as is the case when a regression line is fitted to a scatter diagram consisting of only two points. Thus, the index ρ can serve meaningful purposes only in the restricted cases, i.e., when the number of links is greater than m - 1.

PART C

NUMERICAL EVALUATION

CHAPTER V

EVALUATION DESIGNS

Overview

In order to evaluate the new equating method proposed in Part B, it was applied to computer-generated data sets representing the closed chain design. (For applications of the unrestricted case of the method to real data, see Haebara, 1980; Shiba, Ohama, & Noguchi, 1981; and Albanese, 1981.) The closed chain design, having m links among m test forms, is the simplest of all restrictive designs. An interesting feature of the closed chain design is that the unrestricted solution for this design is equivalent to the solution for the open chain design in which one of the forms in the closed chain is treated as two different forms located at the ends of the open chain. Thus, the comparison between the restricted and the unrestricted solutions for the closed chain design may be regarded as the comparison between the closed and the open chain designs.

Three types of evaluation were performed to investigate different aspects of the method. In Evaluation 1, 100 data sets were independently generated for the four-form closed chain design (see Figure 2 in Chapter II), and the accuracy and some other properties of the estimates of equating constants obtained by the method were investigated. Both restricted and unrestricted solutions were examined.

In Evaluation 2, the estimates of equating constants obtained in Evaluation 1 were used to place all item response functions on the common scale and a common item response function was estimated for each of the pairwise common items. Then the parameters of these estimated common item response functions were compared with the true item parameters, and the errors were evaluated in relation to the errors of estimation to be expected from the joint estimation of all item parameters in all test forms.

In Evaluation 3, a data set was generated for each of 27 closed chains which were systematically different in the number of forms involved, in the number of common items in each adjacent pair of forms, and in the hypothetical sample size. Then the method was applied to equate all scales in each data set, using different numbers of comparison points and different weight functions to examine the effect of these factors on the accuracy of the estimated equating constants.

Throughout the numerical evaluation study, the iterative process for estimating equating constants and common item parameters was considered to have converged when the change of the value of every estimate from the preceding iteration was less than or equal to 0.00001. The initial values for the equating constants were computed by using (20), and the initial values for the common item parameters were computed as described in the last section of Chapter III. In the following sections of this chapter, the three evaluation designs are described in detail. The results of the evaluation study are reported and discussed in the next chapter.

Evaluation 1

First, it was assumed that four samples of examinees, each with sample size of 1200, had the means and the SD's of ability parameters on a common scale as shown in Table 4. (The SD's were specified so that the ability distribution in the group with the higher abilities had the larger SD's.) The tabled values correspond to the true equating constants d_j and k_j to be estimated, as will be explained shortly.

Table 4

Means and SD's of Abilities in the 4 Examinee Samples and True Equating Constants

Examinee Sample	Mean (d_j)	SD (k_j)
1	0.0	1.0
2	0.7	1.1
3	0.0	1.0
4	-0.7	0.9

Second, four sets of 30 items each were hypothetically constructed, the parameters of which defined on the common scale are listed in Table 5. The guessing parameter c of each of the 120 items was set to 0.15, a reasonable value for a five-choice item. The difficulty parameters b were equally spaced in the range from -1.50 to 1.45. Then the discriminating powers a were randomly chosen from the six possible values, 0.4, 0.6, 0.8, 1.0, 1.2, and 1.4. The four test forms were then

constructed from these item sets as shown in Table 6. The table also shows the allocation of these test forms to the examinee samples. (The means of abilities in the examinee samples were specified to be about equal to the mean difficulties of the items administered to them.) Thus, the four test forms constitute a closed chain with 30 items shared by each pair of adjacent forms.

Third, the original scale was hypothetically derived for each test form and each examinee sample in such a way that the mean and the SD of the abilities in the examinee sample were 0 and 1, respectively, on the corresponding original scale. In other words, each original scale was a standard-score scale [see (75) below]. The item parameters were then transformed accordingly so that they were redefined on the original scales. Since the transformation from the common scale to the j-th original scale was defined by

$$\theta_j = (\theta_{common} - d_j)/k_j, \tag{75}$$

the transformations for the parameters of items in Form j were given by

$$\begin{cases} a_j = k_j a_{common} \\ b_j = (b_{common} - d_j)/k_j \end{cases} \tag{76}$$

[see (2) and (3) in Chapter II].

Fourth, the error-perturbed item parameter estimates were generated by adding a random number to each of the item parameters on the original scales obtained in the preceding step. This was done by assuming that

Table 5

True Item Parameters in the 4 Item Sets Defined on the Common Scale

Item Set	Item No. Across Sets	Item No. Within Set	a	b	c
1	1	1	0.40	-1.50	0.15
	2	2	1.40	-1.45	0.15
	3	3	1.40	-1.40	0.15
	4	4	0.40	-1.35	0.15
	5	5	0.40	-1.30	0.15
	6	6	0.40	-1.25	0.15
	7	7	1.20	-1.20	0.15
	8	8	1.00	-1.15	0.15
	9	9	1.40	-1.10	0.15
	10	10	1.20	-1.05	0.15
	11	11	0.60	-1.00	0.15
	12	12	0.60	-0.95	0.15
	13	13	0.60	-0.90	0.15
	14	14	1.20	-0.85	0.15
	15	15	1.40	-0.80	0.15
	16	16	0.40	-0.75	0.15
	17	17	0.80	-0.70	0.15
	18	18	0.40	-0.65	0.15
	19	19	0.60	-0.60	0.15
	20	20	0.60	-0.55	0.15
	21	21	1.40	-0.50	0.15
	22	22	0.60	-0.45	0.15
	23	23	1.40	-0.40	0.15
	24	24	0.60	-0.35	0.15
	25	25	1.40	-0.30	0.15
	26	26	1.40	-0.25	0.15
	27	27	0.60	-0.20	0.15
	28	28	1.00	-0.15	0.15
	29	29	1.00	-0.10	0.15
	30	30	0.60	-0.05	0.15

Table 5 (Continued)

Item Set	Item No. Across Sets	Item No. Within Set	a	b	c
2	31	1	1.40	0.00	0.15
	32	2	0.80	0.05	0.15
	33	3	1.40	0.10	0.15
	34	4	1.40	0.15	0.15
	35	5	1.40	0.20	0.15
	36	6	0.40	0.25	0.15
	37	7	1.40	0.30	0.15
	38	8	1.40	0.35	0.15
	39	9	0.80	0.40	0.15
	40	10	0.40	0.45	0.15
	41	11	1.20	0.50	0.15
	42	12	1.00	0.55	0.15
	43	13	0.60	0.60	0.15
	44	14	0.60	0.65	0.15
	45	15	1.40	0.70	0.15
	46	16	0.60	0.75	0.15
	47	17	1.00	0.80	0.15
	48	18	1.20	0.85	0.15
	49	19	1.00	0.90	0.15
	50	20	0.80	0.95	0.15
	51	21	0.80	1.00	0.15
	52	22	1.00	1.05	0.15
	53	23	0.60	1.10	0.15
	54	24	1.20	1.15	0.15
	55	25	0.40	1.20	0.15
	56	26	1.00	1.25	0.15
	57	27	0.40	1.30	0.15
	58	28	1.00	1.35	0.15
	59	29	0.60	1.40	0.15
	60	30	1.00	1.45	0.15

Table 5 (Continued)

Item Set	Item No. Across Sets	Item No. Within Set	a	b	c
3	61	1	1.40	1.45	0.15
	62	2	1.00	1.40	0.15
	63	3	0.80	1.35	0.15
	64	4	1.00	1.30	0.15
	65	5	1.00	1.25	0.15
	66	6	0.60	1.20	0.15
	67	7	1.00	1.15	0.15
	68	8	1.00	1.10	0.15
	69	9	1.20	1.05	0.15
	70	10	0.60	1.00	0.15
	71	11	1.20	0.95	0.15
	72	12	0.40	0.90	0.15
	73	13	1.40	0.85	0.15
	74	14	1.20	0.80	0.15
	75	15	1.40	0.75	0.15
	76	16	0.80	0.70	0.15
	77	17	1.00	0.65	0.15
	78	18	0.40	0.60	0.15
	79	19	1.00	0.55	0.15
	80	20	1.40	0.50	0.15
	81	21	0.80	0.45	0.15
	82	22	1.20	0.40	0.15
	83	23	0.60	0.35	0.15
	84	24	0.60	0.30	0.15
	85	25	1.40	0.25	0.15
	86	26	1.00	0.20	0.15
	87	27	1.00	0.15	0.15
	88	28	1.20	0.10	0.15
	89	29	1.40	0.05	0.15
	90	30	0.60	0.00	0.15

Table 5 (Continued)

Item Set	Item No. Across Sets	Item No. Within Set	a	b	c
4	91	1	1.00	-0.05	0.15
	92	2	0.60	-0.10	0.15
	93	3	0.60	-0.15	0.15
	94	4	1.20	-0.20	0.15
	95	5	0.40	-0.25	0.15
	96	6	1.00	-0.30	0.15
	97	7	1.40	-0.35	0.15
	98	8	0.60	-0.40	0.15
	99	9	0.80	-0.45	0.15
	100	10	1.20	-0.50	0.15
	101	11	0.40	-0.55	0.15
	102	12	1.40	-0.60	0.15
	103	13	1.20	-0.65	0.15
	104	14	1.20	-0.70	0.15
	105	15	0.60	-0.75	0.15
	106	16	0.40	-0.80	0.15
	107	17	0.80	-0.85	0.15
	108	18	1.40	-0.90	0.15
	109	19	1.00	-0.95	0.15
	110	20	1.40	-1.00	0.15
	111	21	0.60	-1.05	0.15
	112	22	0.60	-1.10	0.15
	113	23	0.40	-1.15	0.15
	114	24	0.80	-1.20	0.15
	115	25	1.00	-1.25	0.15
	116	26	1.20	-1.30	0.15
	117	27	1.20	-1.35	0.15
	118	28	0.40	-1.40	0.15
	119	29	1.20	-1.45	0.15
	120	30	0.80	-1.50	0.15

Table 6

Organization of the 4 Test Forms and Their Allocation to Examinee Samples

Form	Item Sets	Sample
1	1 and 2	1
2	2 and 3	2
3	3 and 4	3
4	4 and 1	4

the sample estimator of an item parameter was independently and normally distributed with mean equal to its true value and variance equal to the reciprocal of the information function evaluated at the true parameter value. The formulas for the variances are given in Haebara (1979, p. 16) and Lord (1980, p. 220). It is known that the maximum likelihood estimator of an item parameter has an asymptotic distribution identical to the one assumed here if the ability parameters and the other parameters of the item are fixed at their true values (see Haebara, 1979, pp. 15-16; see also Cramér, 1946, pp. 500-506; Kendall & Stuart, 1973, pp. 45-46).

In using these formulas, it was hypothesized that the ability parameters in each sample assumed 61 equally spaced values in the range between -3.0 and 3.0 with the frequencies proportional to the standard normal probability density function at these points. The sum of the frequencies was set to 1200, the assumed sample size. (This procedure yields the sample SD of ability parameters slightly different from

unity, but the consequence of this difference is merely that the accuracy of the generated item parameter estimates is slightly different from what is to be expected from the ability distribution with a unit SD.) The item parameter estimates a and b were then generated, using the IMSL subroutine program GGNML (IMSL, 1980). The guessing parameters were assumed to have been fixed at 0.15.

The fourth step was repeated to obtain 100 independent data sets having the same structure described above. Then the method was applied to equate all original scales to Scale 1, which was chosen as the common scale. Since the inverse transformation of (75) equate all scales to the common scale (Scale 1), the values of k_j and d_j given in Table 4 are the true equating constants to be estimated. In the estimation, 10 equally spaced comparison points were set in the range between -3.0 and 3.0 and the standard normal probability density function was used as the weight function for each original scale.

Both restricted and unrestricted solutions were obtained for each of the 100 data sets, and the index of internal consistency (72) was computed. (Note that there are eight equating constants to be estimated in the unrestricted solution, that is, six constants k_2, d_2, k_3, d_3, k_4, and d_4 to be estimated in the restricted solution plus two others, say k_1'' and d_1'', for equating Scale 1 to itself through Scales 2, 3, and 4. The true values of k_1'' and d_1'' are 1 and 0, respectively.) The estimates of the equating constants were then compared with the true values, and the magnitudes of the estimation errors were evaluated. In addition, the correlations among the estimates were examined.

Evaluation 2

For each of the 100 data sets generated in Evaluation 1, a common item response function for each of the pairwise common items was estimated on the common scale by the procedure described in the last section of Chapter III. The estimated item parameters were then compared with the true parameters in Table 5 and the accuracy of estimation was examined. In particular, approximate standard errors and mean deviations of item parameter estimates to be expected from the joint estimation of the parameters of all items in all test forms were computed, and the accuracy of the item parameter estimates obtained by the proposed method was evaluated against these values.

The approximate standard errors just mentioned were computed for common item g shared by Form j' and Form j as follows. First, the information function was computed for each item parameter on Scale j' and Scale j, separately, as was done in Evaluation 1. Next, the weighted sum of the values of the information function on the two scales was taken, with the weights being the squares of $k_{j'}$ and k_j for the parameter a and the squares of $1/k_{j'}$ and $1/k_j$ for the parameter b. Then the square root of the reciprocal of the resulting weighted sum was taken to be the approximate standard error of the jointly estimated item parameter. As mentioned earlier, the guessing parameter was held fixed throughout the analysis.

In computing the standard errors in this way, maximum likelihood estimation was implied again. Taking the weighted sum of the values of the information function can be justified by the additive property of

the information function, and the particular weights used were derived by taking into account the transformations performed on the item parameter estimates on the original scales.

The corresponding mean deviation (the expected value of the absolute deviation of the estimate from the true value) was computed from the approximate standard error by multiplying it by the square root of $2/\pi$. This relationship between the standard error and the mean deviation can be proved if a normal distribution is assumed for the estimator of each item parameter (Kendall & Stuart, 1977, p. 146).

Evaluation 3

In Evaluation 1 and Evaluation 2, only the four-form closed chain design with 30 common items for each adjacent pair of forms and with sample size of 1200 in each sample was used. In Evaluation 3, the number of forms (m) in the closed chain, the number of common items (r) in each adjacent pair of forms, and the sample size (N) were systematically varied, and a total of 27 different closed chains were analyzed. Each of the three factors, m, r, and N, had three levels: m = 4, 8, 12; r = 10, 20, 30; N = 400, 800, 1200. And these levels were completely crossed to yield 3 x 3 x 3 = 27 different combinations of the levels.

The procedure used in Evaluation 1 was followed to generate these different closed chains. For m = 4, the means and the SD's of ability parameters in the samples were the same as those listed in Table 4. The means and the SD's for m = 8 and m = 12 are shown in Tables 7 and 8,

respectively. (Here, again, the SD's of the groups with high mean abilities were assumed to be larger than the SD's of the groups with low mean abilities.) These values are the true equating constants to be estimated.

Table 7

Means and SD's of Abilities in the 8 Examinee Samples and True Equating Constants

Examinee Sample	Mean (d_j)	SD (k_j)
1	0.0	1.0
2	1.5	1.2
3	2.2	1.3
4	1.5	1.2
5	0.0	1.0
6	-1.5	0.9
7	-2.2	0.8
8	-1.5	0.9

The true item parameters in the 12 item sets used for m = 12 and r = 30 were specified on the common scale as follows. First, as in Evaluation 1, the guessing parameters were fixed at 0.15 for all items, and the discriminating powers were randomly chosen from the six possible values, 0.4, 0.6, 0.8, 1.0, 1.2, and 1.4. Then the item difficulties were equally spaced in the ranges shown in Table 9. The increment and the decrement were 0.05; for example, the item difficulties of the first five items in Item Set 1 were -1.50, -1.45, -1.40, -1.35, and -1.30.

Table 8

Means and SD's of Abilities in the
12 Examinee Samples and
True Equating Constants

Examinee Sample	Mean (d_j)	SD (k_j)
1	0.0	1.0
2	1.5	1.2
3	3.0	1.4
4	3.7	1.5
5	3.0	1.4
6	1.5	1.2
7	0.0	1.0
8	-1.5	0.9
9	-3.0	0.8
10	-3.7	0.7
11	-3.0	0.8
12	-1.5	0.9

Then, Form j (j = 1, 2,..., 12) was constructed by combining Item Sets j and j + 1, where j + 1 was taken to be 1 if j = 12. For m = 8, Item Sets 1, 2, 3, 6, 7, 8, 9, and 12 were used to construct the 8 forms by successively combining the adjacent pairs of item sets; for example, Form 2 consisted of Item Sets 2 and 3, and Form 3 consisted of Item Sets 3 and 6. For m = 4, Item Sets 1, 2, 7, and 8 were used and the test forms were constructed similarly. (For m = 4 and r = 30, the items in the test forms were the same items as used in Evaluation 1 and Evaluation 2. For all three levels of m, the means of abilities in the examinee samples were specified to be about equal to the mean difficulties of the items administered to them.)

Table 9

Ranges of Item Difficulties in the
12 Item Sets Defined on the Common Scale

Item Set	Number of Items	Item Difficulties	
		From	To
1	30	-1.50	-0.05
2	30	0.00	1.45
3	30	1.50	2.95
4	30	3.00	4.45
5	30	4.45	3.00
6	30	2.95	1.50
7	30	1.45	0.00
8	30	-0.05	-1.50
9	30	-1.55	-3.00
10	30	-3.05	-4.50
11	30	-4.50	-3.05
12	30	-3.00	-1.55

In each case, every third item, starting with Item 2, was chosen from each item set for r = 10, and the remaining 20 items in each item set constituted the item sets for r = 20. Finally, the three different sample sizes were imposed by setting the sum of the frequencies of the 61 equally spaced ability parameters to the specified sample size (N = 400, 800, or 1200) in generating the error-perturbed item parameter estimates as was done in Evaluation 1.

Each of the 27 data sets constructed was analyzed and the equating constants were estimated in four different ways. First, the restricted solution was obtained with 10 equally spaced comparison points (n = 10) in the range between -3.0 and 3.0 and with the standard normal probability density function used as the weight function for each original scale. Second, the unrestricted solution was obtained under the same conditions regarding the comparison points and the weights. Third, the restricted solution was obtained with 20 equally spaced comparison points (n = 20) in the same range and with the normal weight functions. Fourth, the restricted solution was obtained with 10 comparison points and with the uniform weight functions in the same range.

Then the second, third, and fourth results were compared with the first result to examine, respectively, the effect of the restrictions for internal consistency, the effect of the number of comparison points, and the effect of different weight functions. In addition, the results for the 27 different closed chains were compared to investigate the effect of the number of forms, the number of common items, and the sample size on the accuracy of the estimation of equating constants.

Standardized Criterion Function

In evaluating and comparing the results of equating constant estimation for different designs or under different conditions, the criterion function T is not very convenient because it changes as a function of the number of forms, common items, and comparison points and also as a function of the magnitude of the weights. This difficulty can be avoided by defining the standardized criterion function, say S, by

$$S = (T/w)^{1/2}, \tag{77}$$

where w is the sum of all weights, that is,

$$w = \underline{1}'\underline{W}\underline{1}, \tag{78}$$

where $\underline{1}$ is a column vector containing 1's.

The minimized value of S then represents the root mean squared error and can be interpreted as the average amount of deviation of the obtained equating result from the perfect equating, with the deviation being measured on the probability scale. That is, min(S) indicates, on the average, how much the corresponding item response functions (i.e., the item response functions of the common items) differ after all item response functions are placed on the common scale. The standardized criterion function S* for the unrestricted solution can be defined similarly by dividing T* by w. Then the index of internal consistency of the equating system defined in (72) can be expressed as

$$\rho = [\min(S^*)/\min(S)]^2. \tag{79}$$

CHAPTER VI

RESULTS AND DISCUSSION

Evaluation 1

Estimation of the 6 equating constants in the restricted solution and of the 8 equating constants in the unrestricted solution for the four-form closed chain design took 3 to 5 iterations for convergence. Tables 10 and 11 show the means and the SD's of the 100 estimates of each equating constant for the restricted and the unrestricted solutions, respectively. The tables also show the estimation bias for each equating constant which was defined by the deviation of the mean of the 100 estimates from the true value given in Table 4.

It is clear from the tables that the restricted solution is consistently more accurate than the unrestricted solution as expected; the SD's and the biases are smaller for the restricted solution, with the only exception that the bias for k_4 is smaller for the unrestricted solution. However, even the largest bias in the unrestricted solution is only 0.0070, indicating that both solutions provide reasonably unbiased estimates. Since the unrestricted solution in the closed chain design may be regarded as the solution for the open chain design, the result implies that, at least when the number of forms is as small as 5 (= 4 + 1) and the number of common items is as large as

Table 10

Means and SD's of the 100 Estimates of Equating Constants: Restricted Solution

Equating Constant	Mean	SD	Bias
k_2	1.1026	0.0192	0.0026
k_3	0.9996	0.0158	-0.0004
k_4	0.8976	0.0130	-0.0024
d_2	0.7003	0.0149	0.0003
d_3	-0.0017	0.0133	-0.0017
d_4	-0.7026	0.0127	-0.0026

30, the method yields a reasonably accurate solution to the open chain design.

The difference between the two solutions in accuracy was also examined in terms of the average absolute error, i.e., the mean of the absolute deviations of the estimates from the true value for each equating constant. In the restricted solution, the mean of these average absolute errors was 0.0128 for k's and 0.0107 for d's. The corresponding values in the unrestricted solution were 0.0252 and 0.0258. Thus, in terms of the average absolute error, the restricted solution was found to be about twice as accurate as the unrestricted solution.

Table 11

Means and SD's of the 100 Estimates of Equating Constants: Unrestricted Solution

Equating Constant	Mean	SD	Bias
k_1''	1.0070	0.0409	0.0070
k_2	1.1040	0.0246	0.0040
k_3	1.0022	0.0307	0.0022
k_4	0.9018	0.0319	0.0018
d_1''	0.0020	0.0372	0.0020
d_2	0.7007	0.0164	0.0007
d_3	-0.0028	0.0276	-0.0028
d_4	-0.7054	0.0473	-0.0054

The minimized values of the standardized criterion function S defined in (77) were averaged over the 100 estimates for each solution. (In this chapter, the symbolic distinction between S and S* is not made because the distinction is obvious from the context.) The mean of the minimized S was 0.0241 for the restricted solution and 0.0240 for the unrestricted solution. The closeness of these two was expected because of the way in which the data were generated. That is, the item parameter estimates were randomly generated for given true parameters defined on the common scale, and there was no intrinsic inconsistency among the derived original scales such as the shift of underlying dimensions. The slight reduction in the minimized value of S for the unrestricted solution was, therefore, purely the result of this solution capitalizing on random errors to a greater extent. The mean of the internal consistency index ρ for the 100 data sets was 0.9919, and the SD was 0.0077.

As for the inter-estimate correlations, some high correlations were found among the estimates of k's and among the estimates of d's. To see how well the factors representing the cluster of k's and the cluster of d's can be identified, the two-factor solution obtained by the principal factor method with iteration was transformed to be maximally congruent with a target solution. (There were only two eigen values of the correlation matrix greater than 1 in both restricted and unrestricted cases.) The target solution was represented by a hypothetical factor pattern matrix containing 0's and 1's, the 1's indicating the classification of the variables (the estimates of the equating

constants) to the clusters (see Tables 12 and 13 below). The transformation of the initial factor pattern was performed by using an oblique Procrustes method originally developed by Mosier (1939) which was rediscovered by Korth and Tucker (1976) as the method to obtain a maximum congruence with a target matrix. The results are presented in Tables 12 and 13 for the restricted and the unrestricted solutions, respectively. The parenthesized values are the elements of the target matrix.

Table 12

Oblique Procrustes Factor Solution for the Correlations among the Estimates: Restricted Solution

Equating Constant	Factor I	Factor II	Communality
k_2	0.6302 (1)	0.0147 (0)	0.4006
k_3	0.9458 (1)	0.0205 (0)	0.9018
k_4	0.4712 (1)	-0.0302 (0)	0.2180
d_2	0.5449 (0)	0.5543 (1)	0.7106
d_3	-0.0706 (0)	0.9151 (1)	0.8197
d_4	-0.4675 (0)	0.5282 (1)	0.4106

In both solutions, marked differences between the transformed factor pattern and the hypothesized target matrix are observed in the loadings on the first factor; the second factor contributes only to the

Table 13

Oblique Procrustes Factor Solution for the Correlations among the Estimates: Unrestricted Solution

Equating Constant	Factor I	Factor II	Communality
k_1''	0.8772 (1)	0.0272 (0)	0.7481
k_2	0.7602 (1)	0.0091 (0)	0.5716
k_3	0.9264 (1)	-0.0907 (0)	0.9440
k_4	0.9168 (1)	-0.0223 (0)	0.8598
d_1''	0.0322 (0)	0.8611 (1)	0.7169
d_2	0.5355 (0)	0.8247 (1)	0.5591
d_3	-0.1648 (0)	0.8955 (1)	0.9653
d_4	-0.5000 (0)	0.6558 (1)	0.9823

estimates of d's, but the first factor contributes not only to the estimates of k's but also, to a lesser extent, to the estimates of d's. To put it in another way, the estimates of k's can be adequately represented by a single factor (Factor I), but the estimates of d's cannot. Since the correlations among the estimates are equivalent to the correlations among the estimation errors, these results suggest that overestimation (underestimation) of some k's is associated with overestimation (underestimation) of other k's. But no systematic pattern is found in the correlations among the estimation errors for d's or in the correlations between the errors for k's and the errors for d's. (The inter-factor correlation was 0.1761 for the restricted solution and -0.4617 for the unrestricted solution.)

Evaluation 2

Basic results of Evaluation 2 are summarized in Table 14. The bias was defined for each item parameter by the mean of the deviations of the 100 item parameter estimates from the true parameter value given in Table 5. The means of these biases over the 120 items are given in Row 1 of Table 14. The mean of the average absolute errors in Row 2 is a similar statistic, which was computed by taking the average of the absolute deviations of the 100 estimates from the true value and then computing the mean of these averages over the 120 items. The statistics in Row 3 are the means, over items, of the SD's of the 100 estimates of the respective item parameters. Rows 4 and 5 show the means, over items, of the ratios of the average absolute errors to certain indices

of errors of estimation expected from the joint estimation of the parameters of all items in all test forms. The approximate standard errors and the approximate mean deviations used as the denominators in Rows 4 and 5, respectively, were computed as described in Chapter V.

The tabled values indicate again that the restricted solution is cosistently superior to the unrestricted solution. Of special importance in this part of the evaluation study, however, is examination of the accuracy of the item parameter estimation provided by the proposed method in comparison with the accuracy expected from the joint estimation of all item parameters in the entire data set. As mentioned earlier, the maximum likelihood method is assumed for the item parameter estimation both in the joint estimation and in the separate estimation prior to the application of the proposed method.

The statistics shown in Rows 4 and 5 in Table 14 are relevant to this comparison. Of particular importance are the means of the ratios of the absolute errors of the item parameter estimates obtained by the proposed method to the approximate mean deviations of the estimates of the corresponding item parameters expected from the joint estimation, which are given in Row 5. The values of this statistic are about 1.06 for both the parameters a and b in the restricted solution. Therefore, the restricted solution in the closed chain design (of the size being used, at least) yields almost as accurate item parameter estimates as the joint estimation does. The accuracy of the unrestricted solution is lower than that of the restricted solution, but not grossly so. The ratios in this case are about 1.68 and 1.84 for the parameters a and b, respectively.

Table 14

Various Comparisons Between the Restricted and the Unrestricted Solutions in Item Parameter Estimation

Summary Statistics	Restricted Solution		Unrestricted Solution	
	Item Parameters			
	a	b	a	b
(1) Mean Bias	-0.0020	-0.0007	-0.0022	-0.0019
(2) Mean of Average Absolute Errors	0.0434	0.0379	0.0613	0.0592
(3) Mean of SD's of Parameter Estimates	0.0542	0.0473	0.0765	0.0735
(4) Mean of Ratios of Average Absolute Errors to Approximate Standard Errors	0.8431	0.8467	1.3424	1.4720
(5) Mean of Ratios of Average Absolute Errors to Approximate Mean Deviations	1.0567	1.0611	1.6825	1.8449

Thus, when separately estimated sets of item parameters are available, the proposed equating method may be applied to these estimates to obtain the parameter estimates of all items on a common scale, instead of ignoring the available information and re-estimating all item parameters jointly. Then the obtained estimates would be almost as accurate as the jointly re-estimated values, and computation would surely be much less expensive than the joint re-estimation. Furthermore, inferring from the present results, the accuracy of the item parameter estimates obtained by the present method tends to become even closer to the accuracy gained in the joint estimation as the interconnection among test forms or test levels increases, e.g., from the open chain to the closed chain and to the complete network.

The item parameter estimation procedure that consists of separate estimation of item parameters in individual test forms, transformation of the obtained estimates, and estimation of the common item parameters of each common item has a certain advantage over the joint estimation not only when the separately estimated item parameters are available but also when they are not available. It is often the case that the assumption of unidimensionality, which implies content homogeneity, is reasonably well satisfied within individual test forms, but not for the collection of these forms. In such a case, item response theory may still be applied to analyze and improve each test form separately. Thus, it seems appropriate, when there are several test forms, to start with estimating item parameters and examining the degree of model fit for each test form separately. Then, if the model fit is reasonably

good for each form, the obtained scales may be equated and the common item parameters may be estimated. At this stage, the index of internal consistency can be computed if the equating design is a restrictive design, and the degree of model fit can be judged on the basis of this index. If the equating system is found to be internally consistent, the obtained item parameter estimates on the common scale can be used for further analysis of the collection of test forms; if not, the original item parameter estimates for individual test forms may still be used for analysis of each form.

Evaluation 3

Among the various factors involved in the equating constant estimation in Evaluation 3, only the number of common items (r) in each adjacent pair of forms affected the number of iterations needed for convergence. For each r, a total of 36 solutions were obtained (3 different numbers of forms x 3 different sample sizes x 4 different estimation conditions), and the range of the number of iterations was 8 to 13 for r = 10, 9 to 12 for r = 20, and 4 to 6 for r = 30. Thus, a large reduction in the number of iterations was observed between r = 20 and r = 30.

Main results of Evaluation 3 are shown in Tables 15, 16, and 17. For each combination of the number of forms (m), the number of common items (r), the sample size (N), and the condition of estimation, Table 15 reports the mean of the absolute errors of the estimates of k's. That is, each cell represents the result of a single analysis, but the

tabled values are the means computed from the estimates of (m - 1) k's for Conditions 1, 3, and 4, and of m k's for Condition 2. Similarly, Table 16 shows the means of the absolute errors of the estimates of d's. The minimized values of the standardized criterion function are given in Table 17. Some marginal means computed from these three tables are given in Table 18 to be presented later in this section.

Let us first examine the effect of four different estimation conditions on the accuracy of estimation and the minimized value of S. By using Condition 1 (the restricted solution with n = 10 and with the normal weights) as the basic condition, the results obtained under the other three conditions can be compared with the results obtained under Condition 1. Although there is a substantial amount of irregularities due to the fact that each tabled number is based on only a single analysis, some important facts can still be observed. As for the accuracy of the estimation of k's shown in Table 15, the unrestricted solution (Condition 2) is generally less accurate than the restricted solution. This is true especially for a large number of forms (m) and a small number of common items (r), e.g., for m = 8 and r = 10, m = 8 and r = 20, and m = 12 for all levels of r. Large differences are also found for m = 4 and r = 20. No improvement in the accuracy of estimation of k's is observed when the number of comparison points (n) is increased from 10 to 20 (Condition 1 vs. Condition 3), and no systematic difference is found for different weights (Condition 1 vs. Condition 4).

Table 15

Means of Absolute Errors of Estimates of k's
in 27 Closed Chains under 4 Conditions

m	r	N	Condition 1	Condition 2 (Unrestricted)	Condition 3 (n = 20)	Condition 4 (Uniform)
4	10	400	0.0820	0.2733	0.0829	0.1261
4	10	800	0.2534	0.2193	0.2562	0.2691
4	10	1200	0.2660	0.2129	0.2688	0.2872
4	20	400	0.0918	0.3466	0.0923	0.0711
4	20	800	0.0610	0.3309	0.0617	0.0359
4	20	1200	0.0280	0.3164	0.0282	0.0312
4	30	400	0.0304	0.0270	0.0300	0.0240
4	30	800	0.0207	0.0259	0.0206	0.0176
4	30	1200	0.0038	0.0214	0.0037	0.0069
8	10	400	0.2170	0.5246	0.2175	0.2248
8	10	800	0.2260	0.4714	0.2305	0.2341
8	10	1200	0.2134	0.4328	0.2177	0.1939
8	20	400	0.0756	0.4082	0.0763	0.0884
8	20	800	0.0624	0.3390	0.0635	0.0726
8	20	1200	0.0494	0.3512	0.0508	0.0722
8	30	400	0.0159	0.0130	0.0153	0.0188
8	30	800	0.0187	0.0278	0.0188	0.0197
8	30	1200	0.0094	0.0313	0.0092	0.0115
12	10	400	0.1923	0.5208	0.1998	0.2531
12	10	800	0.1762	0.5434	0.1839	0.2022
12	10	1200	0.1419	0.5426	0.1404	0.1527
12	20	400	0.0718	0.4513	0.0706	0.1128
12	20	800	0.0743	0.3743	0.0741	0.1082
12	20	1200	0.0823	0.3896	0.0823	0.0985
12	30	400	0.0388	0.2850	0.0393	0.0438
12	30	800	0.0275	0.0768	0.0268	0.0217
12	30	1200	0.0088	0.0227	0.0086	0.0155

Table 16

Means of Absolute Errors of Estimates of d's
in 27 Closed Chains under 4 Conditions

m	r	N	Condition 1	Condition 2 (Unrestricted)	Condition 3 (n = 20)	Condition 4 (Uniform)
4	10	400	1.0326	0.8308	1.0333	1.0518
4	10	800	1.0995	0.8690	1.1007	1.0878
4	10	1200	1.1215	0.8417	1.1228	1.1174
4	20	400	0.4659	0.3750	0.4659	0.4761
4	20	800	0.4760	0.3725	0.4762	0.4856
4	20	1200	0.4934	0.3882	0.4934	0.4991
4	30	400	0.0186	0.0268	0.0188	0.0158
4	30	800	0.0165	0.0204	0.0164	0.0146
4	30	1200	0.0062	0.0605	0.0062	0.0089
8	10	400	1.7946	1.4698	1.7965	1.7431
8	10	800	1.7492	1.4537	1.7531	1.7184
8	10	1200	1.7724	1.4621	1.7768	1.7225
8	20	400	0.9204	1.0224	0.9209	0.8844
8	20	800	0.8787	0.9292	0.8784	0.8485
8	20	1200	0.8766	0.9468	0.8767	0.8519
8	30	400	0.0246	0.0265	0.0241	0.0188
8	30	800	0.0327	0.0540	0.0330	0.0385
8	30	1200	0.0135	0.0393	0.0134	0.0194
12	10	400	2.5445	2.2134	2.5450	2.5000
12	10	800	2.5452	2.1935	2.5531	2.4911
12	10	1200	2.5054	2.1928	2.5072	2.4591
12	20	400	1.2337	1.5973	1.2362	1.1558
12	20	800	1.2296	1.5256	1.2294	1.1702
12	20	1200	1.2120	1.5697	1.2108	1.1766
12	30	400	0.0538	0.6039	0.0541	0.0698
12	30	800	0.0344	0.2085	0.0333	0.0287
12	30	1200	0.0206	0.0403	0.0211	0.0357

Table 17

Minimized Values of Standardized Criterion Function in 27 Closed Chains under 4 Conditions

m	r	N	Condition 1	Condition 2 (Unrestricted)	Condition 3 (n = 20)	Condition 4 (Uniform)
4	10	400	0.1357	0.1341	0.1358	0.1067
4	10	800	0.1286	0.1275	0.1286	0.1013
4	10	1200	0.1308	0.1298	0.1308	0.1026
4	20	400	0.1117	0.1098	0.1117	0.0945
4	20	800	0.1108	0.1089	0.1108	0.0944
4	20	1200	0.1093	0.1072	0.1094	0.0911
4	30	400	0.0420	0.0419	0.0420	0.0364
4	30	800	0.0306	0.0305	0.0306	0.0269
4	30	1200	0.0260	0.0256	0.0260	0.0217
8	10	400	0.1197	0.1166	0.1198	0.1000
8	10	800	0.1134	0.1112	0.1135	0.0933
8	10	1200	0.1102	0.1085	0.1102	0.0911
8	20	400	0.1086	0.1073	0.1086	0.0950
8	20	800	0.1095	0.1087	0.1095	0.0972
8	20	1200	0.1056	0.1047	0.1057	0.0936
8	30	400	0.0402	0.0402	0.0402	0.0350
8	30	800	0.0300	0.0299	0.0300	0.0254
8	30	1200	0.0237	0.0237	0.0237	0.0205
12	10	400	0.1108	0.1093	0.1109	0.0934
12	10	800	0.1054	0.1036	0.1054	0.0901
12	10	1200	0.1047	0.1031	0.1048	0.0892
12	20	400	0.1022	0.1008	0.1023	0.0920
12	20	800	0.0964	0.0955	0.0965	0.0859
12	20	1200	0.0974	0.0967	0.0975	0.0884
12	30	400	0.0417	0.0414	0.0417	0.0355
12	30	800	0.0287	0.0287	0.0288	0.0248
12	30	1200	0.0243	0.0242	0.0243	0.0206

As for the accuracy of the estimates of d's shown in Table 16, the unrestricted solution generally does not seem worse than the restricted solution; in fact, the former provides slightly more accurate estimates of d's in many cases. As m or r increases, however, the restricted solution tends to become more accurate. No marked difference is found between Conditions 1 and 3 or between Conditions 1 and 4. The minimized values of S reported in Table 17 are almost identical for r = 10 and r = 20 (Condition 1 vs. Condition 3). The values are only slightly smaller for the unrestricted solution, compared to the restricted solution (the mean internal consistency for the 27 pairs of results was 0.9792). Large differences in the minimized S are found between the results obtained with the normal weights (Conditions 1, 2, and 3) and those obtained with the uniform weights (Condition 4). But the smaller values for Condition 4 do not imply the superiority of the uniform weights. They simply illustrate that the different views on the seriousness of equating errors along the ability scale expressed by the normal distribution and the uniform distribution lead to different amounts of the overall equating errors.

The effect of the number of forms (m) on the accuracy of estimation is not easy to infer, since the true equating constants to be estimated are not the same for different levels of m (see Tables 4, 7, and 8). However, some interaction effects with other factors can still be investigated. As for the accuracy of estimation of k's, an important finding is that the accuracy of the unrestricted solution (Condition 2) decreases more noticeably and consistently than the restricted solution

as m increases. This difference in the pattern of decreasing accuracy is ot clear in the estimation of d's.

The number of common items (r) shared by each adjacent pair of test forms is found to be the most important factor in determining the accuracy of the estimates of equating constants. With only a few exceptions (e.g., r = 10 vs. r = 20 for m = 4 in the unrestricted estimation of k's), there are consistent and substantial improvements in the accuracy of the estimates of both k's and d's as r increases. For the restricted solutions, the means of the absolute errors of the estimates of k's are almost always smaller than 0.1 when r = 20. (The only exceptions are found for m = 12 in Condition 4.) And they are always smaller than 0.05 when r = 30. Similarly, fairly accurate estimates of d's are obtained when r = 20 or larger. For the unrestricted solution, or equivalently, for the open chain design, somewhat larger numbers of common items are needed to achieve the same level of accuracy as the restricted solution.

Finally, as the sample size (N) increases, the estimates tend to be more accurate; however, even increasing N from 400 to 1200 does not provide as much improvement as is provided by increasing the number of common items from 10 to 20 or from 20 to 30. This and some of the other results reported in this section can also be inferred from Table 18, in which the values in Tables 15, 16, and 17 are averaged to show the main effects of the different factors.

Of course, taking the averages of the raw values obscures some important interactions. But general effects of each factor and their

Table 18

Means of the Values in Tables 15, 16, and 17

	Mean of Absolute Errors of Estimates of k's	Mean of Absolute Errors of Estimates of d's	Mean of Minimized Values of S
m = 4	0.1201	0.5002	0.0873
m = 8	0.1478	0.8718	0.0812
m = 12	0.1626	1.2778	0.0763
r = 10	0.2569	1.7159	0.1120
r = 20	0.1443	0.8847	0.1021
r = 30	0.0293	0.0492	0.0308
N = 400	0.1515	0.8963	0.0865
N = 800	0.1457	0.8790	0.0803
N = 1200	0.1334	0.8745	0.0780
Condition 1	0.0940	0.8953	0.0851
Condition 2	0.2807	0.8642	0.0840
Condition 3	0.0952	0.8962	0.0851
Condition 4	0.1042	0.8774	0.0721

magnitudes can be more readily examined in Table 18. Since each value is now based on 36 observations for m, r, and N, and on 27 observations for the estimation conditions, any inference that can be drawn from this table should be more reliable than similar inferences based on the three previous tables. In the table, the largest effect on the accuracy of estimation is observed for the number of common items (r), as expected from the previous results. The means of the absolute errors of the estimates of k's for r = 30 is 11.4 percent of that for r = 10, and the corresponding value for the estimates of d's for r = 30 is only 2.9 percent of that for r = 10. Another important result, which was mentioned above and can be confirmed by Table 18, is that the sample size (N) has only a minor influence on the accuracy of the estimates, compared with the effect of the number of common items. It can also be verified that increasing the number of comparison points from 10 to 20 does not improve the accuracy (Condition 1 vs. Condition 3).

CHAPTER VII

SUMMARY

General Equating Method

In item response models, the unit and the origin of the ability scale are arbitrary, and they are typically determined on the basis of sample statistics in the ability and item parameter estimation process. Therefore, if several ability scales have been derived separately from different test forms (or test levels) administered to different samples of examinees, these scales must be equated to a common scale so that all ability parameter estimates and all item parameter estimates can be directly interpreted and compared on the common scale.

In this paper, a general method for equating logistic ability scales that uses a restricted or unrestricted least squares estimation procedure was developed. The method is applicable when the test forms are linked by common items. Each common item may be shared by more than two test forms, and any linkage pattern can be used (see Figures 1 through 7). When there are more than m - 1 links among m test forms, the method provides a restricted solution to insure an internally consistent system of equatings.

The method is characterized by the explicit definition of the equating error on the probability scale, the minimization of the overall

equating error, and the coherence of the equating in all directions. The researcher can use appropriate weight functions in the minimization in order to reduce the equating errors in the ranges of the ability scale where the errors are considered to be more serious. An index of internal consistency ρ introduced in the paper is useful in judging the consistency among the separately derived ability scales and evaluating the goodness of overall model fit.

Numerical Evaluation of the Method

To evaluate the accuracy of the solution obtained by the proposed method, many artificial data sets with specified characteristics were generated and the method was applied to them. Among the various equating designs that can be used with the method, the closed chain design (see Figure 2) was selected for the evaluation study because of its convenient property that the unrestricted solution for this design with m test forms may be regarded as the solution for the open chain design (see Figure 1) with m + 1 forms. Since the true values of the equating constants and of the item parameters were known, the obtained estimates were compared with them to determine the accuracy of the solution.

The results of the application of the method to 100 randomly generated four-form closed chains with 30 common items in each pair of linked test forms indicated that the estimates of the equating constants were reasonably unbiased. (The hypothetical sample size was 1200, and 10 comparison points were used for each scale in the estimation.) In

terms of the average absolute error, the restricted solution was found to be about twice as accurate as the unrestricted solution. The accuracy of the parameter estimates of each common item obtained by the restricted estimation was very close to what would be expected from the joint estimation of the parameters of all items in all test forms.

The analysis of 27 computer-generated closed chains with different characteristics clearly showed that the number of common items shared by each pair of linked forms was the most important factor in determining the accuracy of the estimates of the equating constants. When the number of items was 30 and the sample size was as large as 1200, the accuracy of the estimates was satisfactory both in the restricted and the unrestricted solutions even when the number of forms was as large as 12.

REFERENCES

Albanese, M. The one-, two- and modified two-parameter latent trait models: An empirical study of relative robustness. Unpublished doctoral dissertation, University of Iowa, 1981.

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Birnbaum, A. Statistical theory of tests of a mental ability. Annals of Mathematical Statistics, 1958, 29, 1285. (Abstract)

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Bock, R. D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 1972, 37, 29-51.

Bock, R. D. Marginal maximum likelihood estimation of item parameters. In H. Wainer (Chair), New developments in item analysis and test scoring. Symposium presented at the meeting of the Psychometric Society, Chapel Hill, N.C., May 1981.

Bock, R. D., & Lieberman, M. Fitting a response model for n dichotomously scored items. Psychometrika, 1970, 35, 179-197.

Cook, L. L., & Eignor, D. R. Score equating and item response theory: Some practical considerations. Paper presented at the meeting of the American Educational Research Association and the National Council on Measurement in Education, Los Angeles, April 1981.

Cramér, H. Mathematical methods of statistics. Princeton, N.J.: Princeton University Press, 1946.

Divgi, D. R. Evaluation of scales for multilevel test batteries. Applied Psychological Measurement, in press.

Douglass, J. B. Applying latent trait theory to a classroom evaluation system: Model comparison and selection. Paper presented at the meeting of the American Educational Research Association, Boston, April 1980. (ERIC Document Reproduction Service No. ED 189 105)

Ferguson, G. A. Item selection by the constant process. Psychometrika, 1942, 7, 19-29.

Haebara, T. A method for investigating item bias using Birnbaum's three-parameter logistic model (ITP Occasional Paper No. 25). Iowa City: University of Iowa, Iowa Testing Programs, December 1979. (ERIC Document Reproduction Service No. ED 185 090)

Haebara, T. Equating logistic ability scales by a weighted least squares method. Japanese Psychological Research, 1980, 22, 144-149.

Haebara, T. Broad-range item calibration by constrained optimization of item parameter transformation. Paper presented at the meeting of the Psychometric Society, Chapel Hill, N.C., May 1981.

Hieronymus, A. N., Lindquist, E. F., & Hoover, H. D. Iowa Tests of Basic Skills, Forms 7/8, Teacher's guide. Boston: Houghton Mifflin, 1978.

Hogg, R. V., & Craig, A. T. Introduction to mathematical statistics (4th ed.). New York: Macmillan, 1978.

International Mathematical & Statistical Libraries, Inc. IMSL library reference manual (8th ed.). Houston: Author, 1980.

Jensema, C. A simple technique for estimating latent trait mental test parameters. Educational and Psychological Measurement, 1976, 36, 705-715.

Jöreskog, K. G. Basic ideas of factor and component analysis. In K. G. Jöreskog & D. Sörbom, Advances in factor analysis and structural equation models. Cambridge, Mass.: Abt Books, 1979.

Kendall, M. G., & Stuart, A. The advanced theory of statistics (Vol. 2, 3rd ed.). London: Griffin, 1973.

Kendall, M. G., & Stuart, A. The advanced theory of statistics (Vol. 1, 4th ed.). London: Griffin, 1977.

Kennedy, W. J., & Gentle, J. E. Statistical computing. New York: Dekker, 1980.

Kolen, M. J. Comparison of traditional and item response theory methods for equating tests. Journal of Educational Measurement, 1981, 18, 1-11.

Korth, B., & Tucker, L. R. Procrustes matching by congruence coefficients. Psychometrika, 1976, 41, 531-535.

Laksana, S., & Coffman, W. E. A comparison of an ANOVA approach and an ICC approach for assessing item bias in an achievement test (ITP Occasional Paper No. 29). Iowa City: University of Iowa, Iowa Testing Programs, September 1980.

Lawley, D. N. On problems connected with item selection and test construction. Proceedings of the Royal Society of Edinburgh, 1943, A61, 273-287.

Lawley, D. N. The factorial analysis of multiple item tests. Proceedings of the Royal Society of Edinburgh, 1944, A62, 74-82.

Lazarsfeld, P. F. The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen, Measurement and Prediction. Princeton, N.J.: Princeton University Press, 1950.

Lazarsfeld, P. F., & Henry, N. W. Latent structure analysis. Boston: Houghton Mifflin, 1968.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. An investigation of item bias in a test of reading comprehension (Technical Report No. 163). Champaign: University of Illinois, Center for the Study of Reading, March 1980. (ERIC Document Reproduction Service No. ED 184 091)

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. Item bias in a test of reading comprehension. Applied Psychological Measurement, 1981, 5, 159-173.

Lord, F. M. A theory of test scores. Psychometric Monograph, 1952, No. 7.

Lord, F. M. An empirical study of item-test regression. Psychometrika, 1965, 30, 373-376. (a)

Lord, F. M. A strong true-score theory, with applications. Psychometrika, 1965, 30, 239-270. (b)

Lord, F. M. An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 1968, 28, 989-1020.

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, N.J.: Lawrence Erlbaum, 1980.

Lord, F. M. Linear transformation of item parameters. Unpublished manuscript, 1981.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Loyd, B. H., & Hoover, H. D. Vertical equating using the Rasch model. Journal of Educational Measurement, 1980, 17, 179-193.

Marco, G. L. Item characteristic curve solutions to three intractable testing problems. Journal of Educational Measurement, 1977, 14, 139-160.

McKinley, R. L., & Reckase, M. D. The fit of ICC's based on two different three-parameter logistic model parameter estimation procedures. Paper presented at the meeting of the Psychometric Society, Chapel Hill, N.C., May 1981.

Mosier, C. I. Determining a simple structure when loadings for certain tests are known. Psychometrika, 1939, 4, 149-162.

Novick, M. R. The axioms and principal results of classical test theory. Journal of Mathematical Psychology, 1966, 3, 1-18.

Novick, M. R., & Jackson, P. H. Statistical methods for educational and psychological research. New York: McGraw-Hill, 1974.

Petersen, N. S., Cook, L. L., & Stocking, M. L. IRT versus conventional equating methods: A comparative study of scale stability. Paper presented at the meeting of the American Educational Research Association, Los Angeles, April 1981.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedagogiske Institut, 1960.

Ree, M. J. Estimating item characteristic curves. Applied Psychological Measurement, 1979, 3, 371-385.

Ree, M. J., & Jensen, H. E. Effects of sample size on linear equating of item characteristic curve parameters. In D. J. Weiss (Ed.), Proceedings of the 1979 computerized adaptive testing conference. Minneapolis: University of Minnesota, Department of Psychology, 1980.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement, 1969, 34(4, Pt. 2).

Samejima, F. A general model for free-response data. Psychometrika Monograph Supplement, 1972, 37(1, Pt. 2).

Samejima, F. A new family of models for the multiple-choice item (Research Report 79-4). Knoxville: University of Tennessee, Department of Psychology, December 1979.

Samejima, F. Research on the multiple-choice test item in Japan: Toward the validation of mathematical models (ONRT M3 Scientific Monograph). Tokyo: Office of Naval Research Tokyo, 1980.

Shiba, S. Construction of a scale for acquisition of word meanings (Bulletin of the Faculty of Education, Vol. 17, pp. 47-58). Tokyo: University of Tokyo, 1978. (In Japanese with English abstract)

Shiba, S., Ohama, K., & Noguchi, H. Verbal ability of Japanese children living abroad (Bulletin of the Faculty of Education, Vol. 20, pp. 111-128). Tokyo: University of Tokyo, 1981. (In Japanese with English abstract)

Swaminathan, H., & Gifford, J. A. Bayesian estimation in item response models. Paper presented at the meeting of the Psychometric Society, Chapel Hill, N.C., May 1981.

Thorndike, R. L., & Hagen, E. Cognitive Abilities Test, Form 3, Examiners manual. Boston: Houghton Mifflin, 1979.

Tucker, L. R. Maximum validity of a test with equivalent items. Psychometrika, 1946, 11, 1-13.

Tucker, L. R. A method for scaling ability test items in difficulty taking item unreliability into account. American Psychologist, 1948, 3, 309-310. (Abstract)

Urry, V. W. Approximations to item parameters of mental test models and their uses. Educational and Psychological Measurement, 1974, 34, 253-269.

Urry, V. W. ANCILLES: Item parameter estimation program with normal ogive and logistic three-parameter model options. Washington, D.C.: U.S. Civil Service Commission, Personnel Research and Development Center, 1978.

Wilks, S. S. Scaling and equating College Board tests. Princeton, N.J.: Educational Testing Service, 1961.

Wood, R. L., Wingersky, M. S., & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters (ETS RM 76-6). Princeton, N.J.: Educational Testing Service, 1976.

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

IOWA TESTING PROGRAMS

OCCASIONAL PAPERS

1. Coffman, William E. Moratorium? What Kind? (ED 099 408)

2. Coffman, William E., & Mathews, Walter M. Narrative Reports and the Iowa Tests of Basic Skills. (April 1973)

3. Forsyth, Robert A., & Feldt, Leonard S. On the Validity of the ITED as an Aid in Program Evaluation. (ED 142 558)

4. Gulliksen, Harold O. Applications of Psychological Scaling Methods. (October 1973)

5. Lee, Lai-Min Paul, & Coffman, William E. A Study of the "I Don't Know" Response in Multiple-Choice Tests. (ED 141 371)

6. Novick, Melvin R. A Course in Bayesian Statistics. (Out of print; in American Statistician, May 1975, Vol. 29, No. 2, pp. 94-98)

7. Hood, Joyce, Leslie, Lauren, & Kendall, Janet Ross. Instructional Effectiveness in Primary Grade Reading: A Pilot Study. (March 1974)

8. Petersen, Nancy S., & Novick, Melvin R. An Evaluation of Some Models for Test Bias. (Out of print; in Journal of Educational Measurement, Spring 1976, Vol. 13, No. 1, pp. 3-31)

9. Kendall, Janet Ross, & Hood, Joyce. Instructional Practices and Easy-to-Change Surrounding Conditions Variables in Effective Primary Reading Programs. (ED 123 557)

10. Petersen, Nancy S. An Expected Utility Model for "Optimal" Selection. (Out of print; in Journal of Educational Statistics, Winter 1976, Vol. 1, No. 4, pp. 333-358)

11. Lindley, D. V. The Effect of Ethical Design Considerations on Statistical Analysis. (March 1975)

12. Lindley, D. V. A Class of Utility Functions. (In Annals of Statistics, Vol. 4, No. 1, January 1976)

13. Lindley, D. V. Probability and Medical Diagnosis; and The Role of Utility in Decision-Making. (March 1975)

14. Lindley, D. V. Inference for a Bernoulli Process: A Bayesian View; and The Future of Statistics: A Bayesian 21st Century. (March 1975)

IOWA TESTING PROGRAMS
OCCASIONAL PAPERS - (continued)

15. Shigemasu, Kazuo. Development and Validation of a Simplified M-Group Regression Model. (In Journal of Educational Statistics, Summer 1976, Vol. 1, No. 2, pp. 157-180)

16. Cantor, Gordon N. Sex and Race Effects in the Community Behavior of Upper-Elementary-School-Aged Children. (ED 115 658)

17. Forsyth, Robert A. Describing What Johnny Can Do. (ED 181 063)

18. Sedere, M. U., & Feldt, Leonard S. The Sampling Distribution of the Kristof Reliability Coefficient, the Feldt Coefficient, and Guttman's λ_2. (In Journal of Educational Measurement, Spring 1977, Vol. 14, No. 1, pp. 53-62)

19. Coffman, William E. Those Achievement Tests - What For? (December 1976)

20. Hunt, David E. Teachers are Psychologists, Too: On the Applications of Psychology to Education. (ED 129 651)

21. Johnson, Sylvia T. Self-Concept in a School Setting: Construct Validations by Factor Analysis. (ED 142 573)

22. Coffman, William E., & Shigemasu, Kazuo. Appraising School Effectiveness Using a Bayesian Method. (ED 159 220)

23. Coffman, William E. Classical Test Development Solutions. (ED 175 913)

24. Olsen, Scott A. An Investigation of Possible Test Floor Effects by Comparing Score Distributions for Students Tested with Two Different Levels of a Test Battery at Four Different Times over a Two-Year Interval. (ED 175 914)

25. Haebara, Tomokazu. A Method for Investigating Item Bias Using Birnbaum's Three-Parameter Logistic Model. (ED 185 090)

26. Coffman, William E., & Olsen, Scott A. The First Two Years of PLAN*--An Evaluation of Program Impact. (June 1980)

27. Haebara, Tomokazu. Equating Logistic Ability Scales by a Weighted Least Squares Method. (August 1980) (ED 193 300) (In: Japanese Psychological Research, September 1980, Vol. 22, No. 3, pp. 144-149.)

28. Mayekawa, Shin-ichi, & Haebara, Tomokazu. Estimation of the Reliability of a Test Consisting of More Than Three Congeneric Parts. (August 1980) (ED 193 302)

29. Laksana, Sangob, & Coffman, William E. A Comparison of an ANOVA Approach and an ICC Approach for Assessing Item Bias in an Achievement Test. (September 1980)

IOWA TESTING PROGRAMS
OCCASIONAL PAPERS - (continued)

30. Haebara, T. Least Squares Method for Equating Logistic Ability Scales: A General Approach and Evaluation. (December 1981)